USING STATISTICAL PROCESS CONTROL METHODS TO CLASSIFY PILOT MENTAL WORKLOAD

THESIS

Terence Y. Kudo, Captain, USAF

AFIT/GOR/ENS/01M-10

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U. S. Government.

AFIT/GOR/ENS/01M-10

USING STATISTICAL PROCESS CONTROL METHODS TO CLASSIFY PILOT MENTAL WORKLOAD

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Terence Y. Kudo, B.S.

Captain, USAF

March 2001

AFIT/GOR/ENS/01M-10

USING STATISTICAL PROCESS CONTROL METHODS TO CLASSIFY PILOT MENTAL WORKLOAD

Terence Y. Kudo, B.S.
Captain, USAF

Approved:

Jeffrey W. Lanning (Chairman) — 02 Mar 01 — date

Kenneth W. Bauer (Member) — 02 MAR 01 — date

Acknowledgements

I would first like to thank my faculty advisor, Maj. Jeffrey W. Lanning and my faculty reader, Dr. Kenneth W. Bauer. I would not have made it through this thesis effort without their expertise and guidance. I would also like to thank Dr. Glenn Wilson and Mr. Christopher Russell, from the Air Force Research Laboratory for supporting this endeavor as well.

I would also like to thank all of my classmates, especially those who I had the privilege of working with in the computer lab every day. I'd like to thank my wife who supported me throughout the thesis effort. I'd also like to thank my daughter who always reminded me what's important in life. Last, but certainly not least, I'd like to thank God, for it is through Him all things are possible.

Terence Kudo

Table of Contents

List of Figures

List of Tables

Abstract

The problem of classifying pilot mental workload is important to the United States Air Force. Pilots are more subject to errors and G-induced loss of consciousness during periods of mental overload and task saturation. Often the result is the loss of aircraft, and in extreme cases, the loss of the pilot's life. Current research efforts use different psychophysiological features to classify pilot mental workload. These include cardiac, ocular, respiratory, and brain activity measures.

The focus of this effort is to apply statistical process control methodology on different psychophysiological features in an attempt to classify pilot mental workload. The control charts track these features throughout the flight, and classify a segment as high workload if the measurements of these features are greater than predefined control limits. We find that certain control charts prove to be effective workload classifiers and maintain high classification accuracies when applied to other flight data.

USING STATISTICAL PROCESS CONTROL METHODS

TO CLASSIFY PILOT MENTAL WORKLOAD

1. Introduction

1.1. Statement of the Problem

The purpose of this research is to use statistical process control (SPC) methods to classify pilot mental workload. Past research (East, 2000; Laine, 1999; Greene, 1998) has used artificial neural networks and multivariate discriminant models to classify pilot mental workload using data from both simulated and actual flight. These classifiers attain high levels of classification accuracy when used to classify pilot mental workload for the same pilot on the same day. However, these classifiers do not maintain high classification accuracies when used to classify pilot mental workload on other days or for different pilots. Our initial goal is to discover if SPC methods are suitable ways to classify pilot workload, with a long-range goal of finding a statistical process control method that can accurately classify pilot workload across multiple days for the same pilot, and ultimately across multiple days for different pilots. For the purpose of this research, we use data obtained from two individual pilots, with each pilot flying a prescribed and flight route on two different days.

1.2. Background

The problem of workload has been a concern for many pilots. As technology has increased, the complexity of Air Force weapon systems has increased as well. Due to this complexity, the mental demands on the pilot have increased. Often, pilots have to split their attention among different tasks. When the pilot's focus is divided, the mental stress upon the pilot increases as well. Consequently, these increases in mental demands have caused some pilots

to forget basic flying procedures, such as G-straining maneuvers, and have resulted in several fatalities. Between 1986 and 1995 the USAF lost 14 pilots due to G-induced loss of consciousness (Auten, 1996). Initially, such accidents were attributed to inexperience or pilot error. Further investigation, however, revealed that neither lack of experience nor ability was common among all the accidents. What was common, with only one exception, was that these accidents occurred in the most demanding periods of flight. The conclusion of this investigation was that pilots were becoming mentally overloaded in the cockpit and forgetting their flying fundamentals. In an effort to reduce the risk of mental overload, the Air Force is developing an automated warning system, which would indicate periods of high mental workload. This system could not only protect our pilots, but it could also improve their effectiveness in the cockpit.

1.3. Research Objectives

Previous research has focused on using multivariate discriminant models to classify pilot mental workload. This will be the first attempt to use statistical process control to classify pilot workload. SPC methods track a process and signals when that process has breached predefined control limits. In the case of pilot workload, an automated system could track a psychophysiological feature and signal when that feature exceeds certain limits. The research involves four different tasks. The first task is to identify which psychophysiological features should be used to track using SPC techniques. An obvious first choice would be to use the features found to be important in previous research (Laine, 1999; East, 2000). However, we may look at every psychophysiological feature to determine which ones work best on control charts. The second task is to take the important physiological features and construct a data set, which we can plot on the control charts. Since we are interested in monitoring different features at the same time, we need a method to put them on the same timeline. The third task would be to build these control charts and measure the classification accuracy of each control chart. The final task is to modify these control charts so that we minimize the misidentification and false alarm rates.

Since the development of a warning system for the pilot is the ultimate goal, the classifier must integrate into this larger, more complex system. The hope is that this system will warn the pilot if he or she is in a state of high mental workload. Another hope is that this system will automate tasks when the pilot is in this state of high mental workload. This system must not only be able to accurately classify high mental workload, but it also must not signal high mental workload when the pilot is not mentally overloaded.

1.4. Research Methodology

We must accomplish several tasks if we hope to determine whether SPC methods can accurately classify pilot mental workload. The first task is to determine which of the 151 psychophysiological features should be monitored as a part of this classifier. These input features include brain electric activity, heart rate, respiratory measures, and ocular measures. The next task is to develop data sets to use with statistical classifiers. Since we wish to measure different psychophysiological features at the same time, every feature must be put on the same timeline. The next task is to develop different control charts on the given variables to see if any of them can classify pilot mental workload. We will first use standard control charts, and if standard control charts do not work, then we will investigate if variations on existing control charts can improve the classification accuracy of the control chart. For each control chart classifier, we determine the classification accuracy when used against itself. In other words, is the control chart able to classify pilot mental workload for the same pilot on the same day? Also, we determine the classification accuracy of each classifier across different days and different pilots. Finally, we must compare SPC classifiers against past multivariate discriminant models to see if SPC classifiers provide some kind of improvement over past classifiers.

1.5. Document Overview

This document first covers past research done in the area of classifying pilot mental workload, and the application of SPC techniques to similar problems. It then covers the data processing, initial analysis of the data set, and the use of traditional control charts, which include x-bar charts, EWMA charts, and CUSUM charts. We also modified several of these control charts in the hopes that we would be able to improve the accuracy of the classifier. We conclude this document with our conclusions and recommendations for further research in this area.

2. Background and Literature Review

2.1. Overview

In order to solve the problem of pilot mental workload classification, we must first understand what has been done in the past to solve this problem. This involves understanding the different psychophysiological features that have been used to classify pilot mental workload. This will provide insight as to which features may be the most relevant in classifying pilot mental workload. Once we understand the different psychophysiological features, we can apply SPC methodologies to determine which methods and features produce the best classifiers. We must also know how these methods work, so we know what kinds of modifications can improve the classification accuracy of the classifier.

2.2. Psychophysiological Features

Past research shows that there is some correlation between different psychophysiological features and pilot mental workload. The psychophysiological features that are most commonly measured are cardiac measures, respiratory measures, hormone measures, ocular measures, and brain activity measures. During each test flight, these psychophysiological features were monitored and recorded.

2.2.1. Description of Psychophysiological Features

Past research focused primarily upon correlating pilot mental workload to different psychophysiological features. Recently, research efforts analyzed psychophysiological responses in different multi-task environments (Hanks and Wilson, 1998). Only in recent years have psychophysiological responses been monitored during actual flight. These psychophysiological responses have advantages in different multi-task environments: the

measurements can be collected throughout the study; they can be easily collected during the flight; the features are relatively robust; and the collection does not interfere with the task completion (Wilson, 1998).

2.2.2. Cardiac Measures

Heart rate has been used for many years as a method of measuring mental workload. Past experiments have considered heart rate and heart rate variability (HRV). Heart rate is the number of heartbeats within a given timeframe. In general, heart rate increases as cognitive workload increases (Hanks and Wilson, 1998). This research uses a different cardiac measure, called heart beat interval, which is related to heart rate. This is the time in milliseconds between the current heart beat and the previous heart beat. Heart beat interval decreases as workload increases. Another cardiac measure is the heart rate variability. Heart rate variability is a measure of the variation in the heart rhythm. However, there is some debate as to the correlation between HRV and cognitive workload.

The first problem with HRV is the question of how to measure it. There are numerous ways to measure HRV, and there is no way to know if one method is universally better than another. The utility of a given method may be based solely upon the situation. Therefore, we do not know which type of measurement would be best in our case. Another problem is conflicting conclusions on the advantage of measuring HRV in a given experiment. Some studies conclude that there is a great advantage, while others conclude that there is no significant advantage. The last problem is to determine the usefulness of HRV in relationship to heart rate. Many studies suggest that heart rate is a better predictor of cognitive workload, implying that heart rate alone may be adequate. In general, heart rate variability is thought to decrease as mental workload increases (Wilson, 1992).

2.2.3. *Respiratory Measures*

There have been few studies using respiration as a predictor for mental workload. The respiratory measurements taken for this research are the interval between breaths, and the minimum and maximum amplitude of the breath. In general, the inter-breath interval decreases as workload increases (Damos, 1996). However, taking respiration measurements during tasks that involve verbal communication is very difficult, as they interfere with the breaths. To solve this problem, it has been recommended that voice pattern analysis may be a better measure of cognitive workload than breath interval. Studies show that fatigue and stress can cause measurable changes in the voice pattern (Damos, 1996).

2.2.4. *Hormone Measures*

Another physiological feature is the measurement of hormone levels. High mental workload is often a result of a stressful situation, since these stressful situations require concentration and quick thinking. When the body responds to this stress, the sympathetic nervous system is stimulated. The sympathetic nervous system then changes hormone levels, which attempt to help the body respond to the stress. However, collecting hormone measurements pose a problem, as hormone levels are typically measured in body fluids, such as blood, urine or saliva. It is very difficult to measure these during flight. Therefore, the only feasible way to measure hormone levels is after the task is completed (Damos, 1996). Since the classifier must classify a high mental workload either before or during the stressful situation, a system that classifies workload after the fact will not work for this situation. Therefore, hormone measurements are not practical as a classifier of mental workload.

2.2.5. *Ocular Measures*

Ocular measurements have been collected during flights to predict pilot workload (Hanks and Wilson, 1998). These ocular measurements include the time between eye blinks, the duration of the blink, and the blink amplitude. In general, it has been found that as the visual demands increase, both the inter-blink rate and the blink duration decrease, since both inter-blink rate and blink duration correspond to the amount of time the eyes are closed. If high mental workload is related to increased visual demands, these ocular measurements would be a good classifier. However, research has shown that ocular measurements may only be a good predictor of visual information processing. Mental workload that is not dependent upon visual information (such as remembering checklists and procedures) may not be correlated to ocular measurements at all.

2.2.6. *Brain Activity Measures*

The most exhaustive part of the available data set is the brain electrical activity or electroencephalography (EEG) data. The brain has electric signals running across it. By placing electrodes over the scalp, we can take measurements of these electric signals (East 2000). Brain electrical data is taken at a rate of 256 times a second over 29 different nodes. The data is processed through a Fast Fourier Transform (FFT), which takes the data at each node and separates it into five different frequency bands. In other words, for each second of processed data, the FFT takes 256 lines of raw EEG data, and breaks it down into the appropriate frequency. Therefore, for each of the 29 nodes, there are five different frequencies for a total of 145 separate EEG variables. The different frequencies are shown in Table 2-1.

Table 2-1: Frequency Band Designations, Symbols and Frequency Specifications

Band	Symbol	Frequency
Delta	Δ	1-3 Hz
Theta	θ	4-7 Hz
Alpha	α	8-12 Hz
Beta	β	13-30 Hz
UltraBeta	μβ	31-42 Hz

Brain researchers have found that different frequencies, as well as the node locations, are important in classifying mental workload. For example, the alpha and theta frequencies have been the most useful frequencies in measuring mental workload (East, 2000). Other research has shown that as cognitive demands increase, alpha band activity decreases and theta band activity increases (Hanks and Wilson, 1998; Damos, 1996). Each node is also associated with certain psychophysiological functions. These relationships are shown in Figure 2-1.

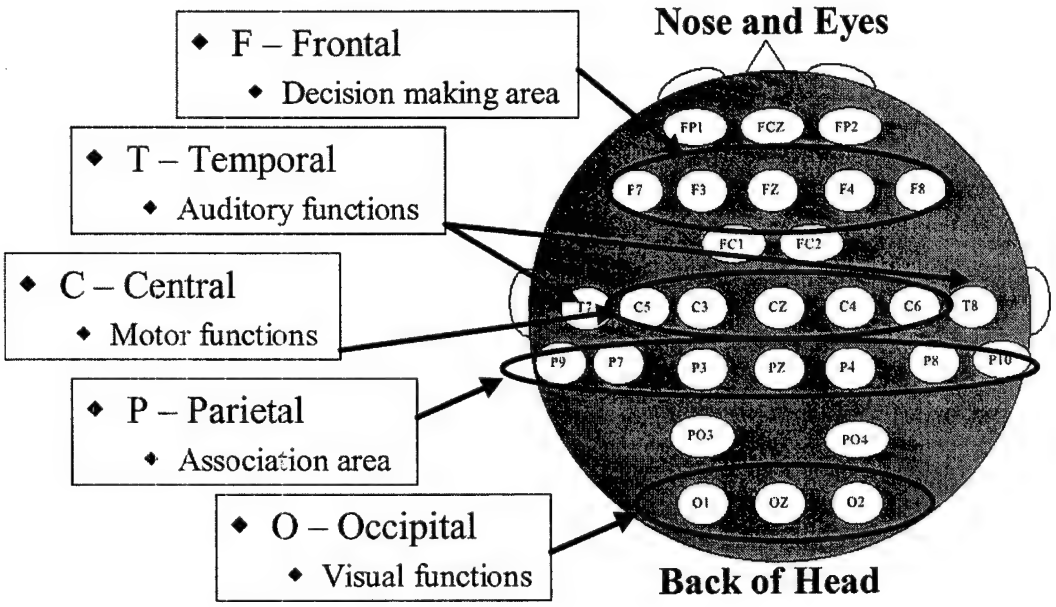

Figure 2-1: Location of Nodes and Relation to Brain Function

2.2.7. *Previous Research*

Past research classifies pilot workload based on several measurable psychophysiological features. In recent years, the Air Force Research Laboratory, Human Effectiveness Directorate (AFRL/HE) at Wright-Patterson AFB, Ohio, has researched this problem (AFRL, 1998). Their research indicates that a handful of psychophysiological features are most influential in classifying pilot mental workload: brain electrical activity, heart rate, breath rate, and eye blink measures (Hankins and Wilson, 1998). The table below shows how these variables respond to an increase in mental workload.

Table 2-2: Response of 8 Different Psychophysiological Features in Response to an Increase in Pilot Mental Workload

Psychophysiological Feature	Feature Response to an Increase in Pilot Mental Workload
Heart Rate	Increase
Heart Beat Interval	Decrease
Heart Rate Variability	Decrease
Number of Eye Blinks	Decrease
Interblink Interval	Increase
Number of Breaths	Increase
Interbreath Interval	Decrease
EEG – Alpha Band	Decrease
EEG – Theta Band	Increase

Initially, 151 psychophysiological features were collected on test subjects in a simulated environment. Laine (1999) used feedforward multiplayer perceptron (MLP) neural networks to classify pilot mental workload using the data from the simulation. Although the resulting classification scheme was found to have promising results, there was no way to know if this classifier could actually classify pilot mental workload. In order to determine if different classification schemes would work against actual flight data, AFRL/HE collected data from ten different pilots flying Wright-Patterson Aero Club Piper Cubs over a predefined route for two days. Each pilot wore special equipment to monitor and record brain electrical data, cardiac,

respiratory, and ocular measurements. Comparing the simulated and the actual data, East (2000) found that a classification scheme could only be valid if it was developed using actual pilot data. She used various screening methods for pilots 1 and 4 and reduced the 151 features, while still maintaining a moderate amount of classification accuracy. Further screening methods reduce the 151 features to 23 for pilot 1 on day 1, while maintaining a classification accuracy between 74% and 84% for same day data. The same screening methods were used on the other three other data sets (pilot 1, day 2, pilot 4, day 1, and pilot 4, day 2) with similar results. However, this classification accuracy diminishes greatly when it used to classify mental workload across days or between different pilots. East (2000) concluded that a classification scheme was still needed that accurately classifies pilot workload across days and pilots. In order to obtain this desired level of accuracy, we will design a classifier using a different technique.

2.3. Quality Improvement

In order to define SPC, we first have to define the process. This type of process improvement is often referred to by such names as "quality improvement" or "total quality management." Montgomery (1997) defines quality improvement as the reduction of variability in processes and products to improve them. It is most often used with manufactured goods as a means to reduce the occurrence of defective products. SPC is a collection of statistical problem solving tools used to achieve process stability and improving capability through the reduction of variability (Montgomery, 1997).

2.3.1. Basic Terminology of Process Improvement

One of the uses of SPC is to identify periods of process instability. Variation inherent in the process if often referred to as "background noise" or a "stable system of chance causes" (Montgomery, 1997). This natural variation exists no matter how well a process is run. When a process shows only this natural variation, it is considered to be "in a state of statistical control."

However, if there is an assignable cause to the variability, we can define the process as being out of control. These causes can be such things as operator error, improper settings or invalid input parameters (Montgomery, 1997). The quality of a process can be improved by monitoring the variability of the output over a period of time. For example, a machine that fills cola bottles will usually put in somewhere between 1998mL and 2003mL of cola. Suppose we find that the machine is not filling the bottles with the right amount of soda. Perhaps the machine is putting in between 1985mL and 2020mL of cola, either not putting in enough cola or overfilling the bottle. We would then conclude that something is wrong with the machine. At this point, we may shut the machine down until we can find the cause of such variation.

2.4. Control Charts

One of the most widely used tools in quality control is the control chart (Montgomery, 1997). A control chart monitors the process and determines the point at which the process is no longer within the given limits of the process. The purpose of a control chart is to determine if the process variation is due to chance or there is an actual change in the process. A shift in the process indicates that the process is out of control.

In a statistical sense, a control chart is a series of hypothesis tests, where each hypothesis test has the null hypothesis "H_o: $\mu=\mu_o$" and the alternate hypothesis "H_a: $\mu\neq\mu_o$." By failing to reject the null hypothesis, we state that $\mu=\mu_o$, or that the process is in control. However, if there is enough evidence to reject the null hypothesis, the process is now out of control. Different types of control charts attempt to prove or disprove this hypothesis in unique ways. An example of a control chart is in Figure 2-2.

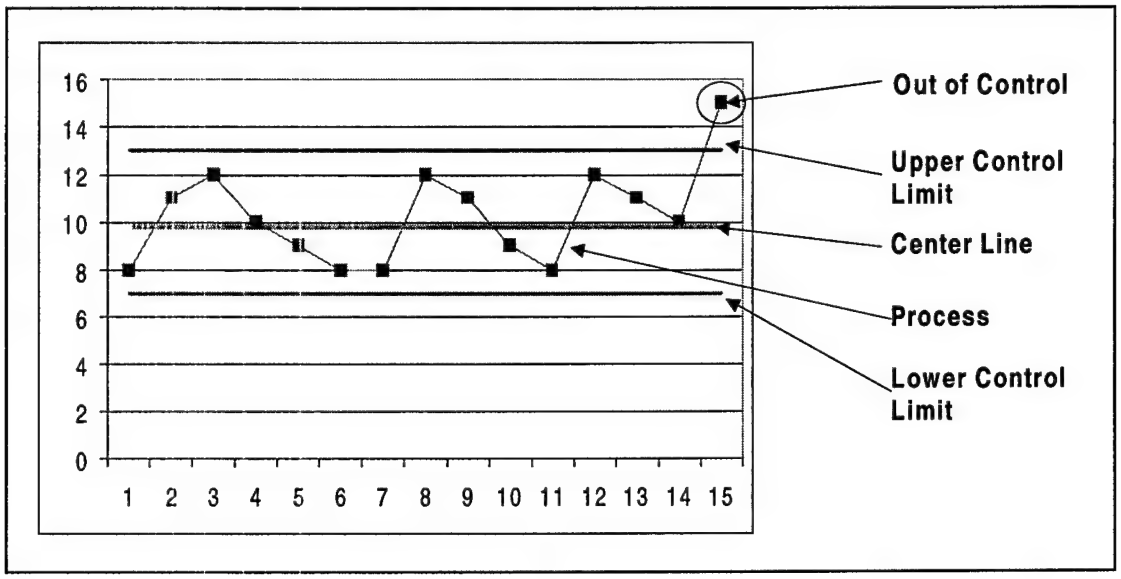

Figure 2-2: Example of a Control Chart

In this chart, the process is considered to be in control while it remains between the two control limits. Once the process goes beyond the control limits, there is a violation, and the process is considered to be out of control. At this point, the control chart stops monitoring the process until the cause of the variation is found.

2.4.1. Using Control Charts as Classifiers

The discussion of control charts has addressed its utility with manufactured goods. However, the purpose of this thesis is to classify pilot mental workload, not to improve a production line. In order to use control charts as a classifier of pilot mental workload, we need to make a couple of modifications to traditional control charts. First, we need to translate what the areas of in control and out of control are. In the case of pilot mental workload, we define the area of in control as low mental workload, and the area of out of control as high mental workload. Since there are only two workload levels, only one control limit is needed. The second modification is that the control chart does not stop monitoring the process when it finds that it is

out of control. Rather, we state that the workload level is high, and continue to monitor the process. This is because we do not want to stop monitoring the process even if we detect high mental workload. An example of what a control chart for pilot mental workload will look like is in Figure 2-3.

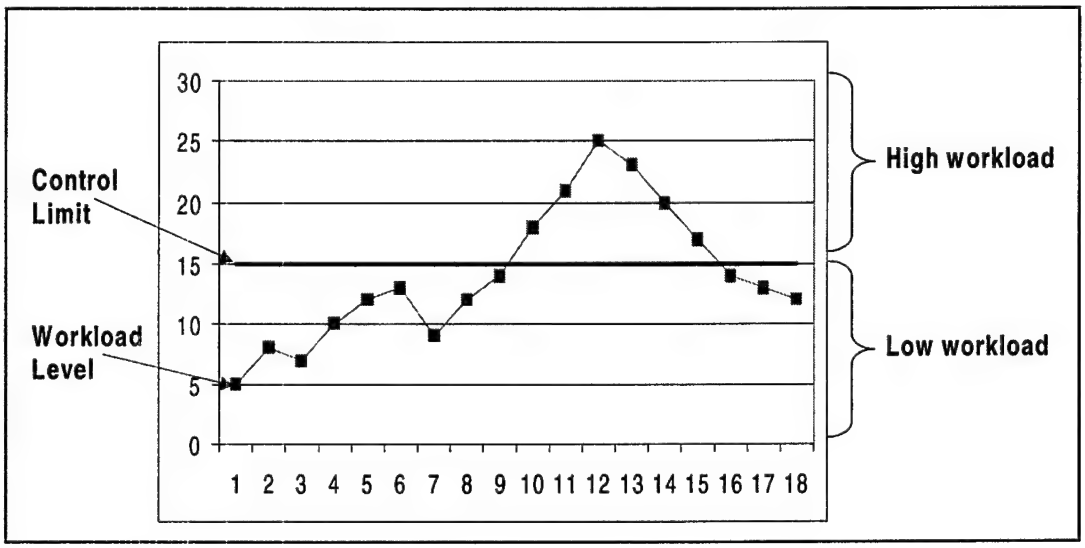

Figure 2-3: Example of a Control Chart as a Classifier of Pilot Mental Workload

In the chart above, there are two regions separated by the control limit. If the workload level is above the control limit, the mental workload is high. If the workload level is below, the mental workload is low.

2.4.2. X-bar Charts

The most common types of control charts used are charts for variables or Shewhart charts (Montgomery, 1997). This type of control chart monitors the mean of each sample, and compares them to predefined control limits to determine if the process is in control or not. To build a chart for variables, we define several statistics relating to the data set. The simplest

measurement we take is the mean (μ), which is simply an average of all the samples in the data set. There will also be a degree of variability in the data, which we measure using the standard deviation (σ). Using these two measurements, we can build the simplest control chart, the x-bar chart. An example of an x-bar chart is shown in Figure 2-4.

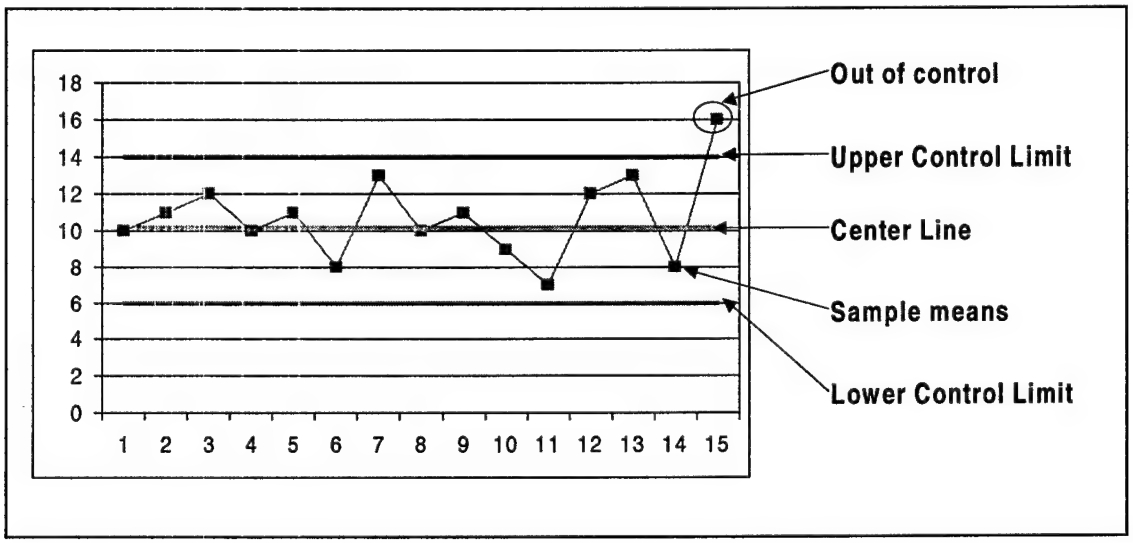

Figure 2-4: Example of an x-bar Chart with Upper and Lower Control Limits

The chart contains a centerline (CL), which is defined by μ, or the average of all the data points. In addition to the center line, there are two control limits, an upper control limit (UCL) and a lower control limit (LCL). These are the points at which the process is defined to be out of control. These three points are defined in Equation 2-1.

$$UCL = \mu + k\sqrt{\sigma^2/n}$$
$$CL = \mu \qquad (2\text{-}1)$$
$$LCL = \mu - k\sqrt{\sigma^2/n}$$

where

n = sample size

k = number of standard deviations between the center line and the control limit

Traditionally, k is assigned to be three, which generates an interval that is six standard deviations wide. The assignment of k determines several characteristics of the x-bar chart. As the value of k increases, the number of data points that fall between these control limits will increase as well. Increasing k will also decrease the chance of accidentally calling the process out of control when the process is actually in control. However, it will also increase the chance of stating that the process is in control when the process is actually out of control. Most control charts use control limits that are three standard deviations separated from the mean in both directions. An x-bar chart with these control limits will on the average capture 99.73% of all data points that are in control. This means that one out of every 370 data points would on the average naturally fall outside of the three standard deviation control limits.

One of the assumptions about an x-bar chart is that the input data is derived from a normal distribution. Therefore, the majority of the data points will be centered about the mean. On the average, a two standard deviation interval centered about the mean (i.e. $\mu \pm \sigma$) will contain 68.26% of the data points. Expanding the length of this interval to four standard deviations will contain 95.44% of the data points, and an interval of six standard deviations will contain 99.73% of the data points. Table 2-3 shows the relationship between the value of k and the probability of in-control data falling within those control limits, and the average number of data points needed before detecting one data point that is out of control.

Table 2-3: Relationship Between Number of Standard Deviations and the Percentage of Data Within Control Limits

Value of k	Percentage of Data Within Control Limits	Average Number of Data Points Before Normally Detecting Out of Control Data
1	68.26%	3.15
2	95.44%	21.93
3	99.73%	370.37
4	99.99366%	15772.87
5	99.9999426%	1742160.28

One of the drawbacks of using an x-bar chart is the assumption of independent data. This means that the x-bar chart only looks at the current sample and does consider historic data. Despite this limitation, there are ways to account for historic data and data trends. One way to account for historic data is to use a method called "run rules." Run rules are a series of rules that use a "run" of data samples to determine if a process is in control. These rules allow the x-bar chart to not only look at the current data point, but also look at historic data points. The Western Electric Company first developed a set of run rules, which include these four rules:

- 1 sample beyond 3 standard deviations
- 2 out of 3 samples in a row beyond 2 standard deviations
- 4 out of 5 samples in a row beyond 1 standard deviation
- 8 samples in a row on the same side of the center line

If any of these four rules are met, then the process is considered to be out of control. These four rules assume an x-bar chart using a standard three-sigma standard deviation. When the criteria for any of these four rules are met, then the process is considered to be out of control. However, Western Electric decided to include four additional rules, which are

- 6 samples in a row either increasing or decreasing
- 15 samples in a row within 1 standard deviation of the center line
- 14 samples in a row alternating up and down
- 8 samples in a row with no sample within 1 standard deviation of the center line

These eight rules make up what are now called the Western Electric run rules. Although they do not take into account all of the previous data points, they do look at some of the historic data to make a decision if the process is in control or not.

The x-bar chart has been used in past research in numerous ways. Air Mobility Command wanted to know if there were an unusually high number of departure delays. Liu (1997) used x-bar charts to monitor the average delay time for departing aircraft over a 13-month period. Using an x-bar chart, they found that a significant number of departures were late. Using this information, they knew to look for assignable causes for these delays, which they then fixed. In another instance, x-bar charts were used to determine a patient's risk of liver disease based upon the patient's work environment. Richardson (1996) used x-bar charts to monitor three different environmental factors that were present in patient's workplaces. He wanted to know if increased levels of these environmental factors would indicate an increased risk of liver disease. He tracked these factors and found that certain levels did increase the patient's risk of liver disease.

2.4.3. R-Charts and S-Charts

Often, it is not a mean shift that affects the stability of a process, but an increase in the variability of a process. A variability increase can affect the process by increasing the number of products that do not meet the required specification. One method of tracking the process variability is the R-chart. An R-chart tracks the range of each sample and signals an out of control process when the range is too large. An example of an R-chart is shown in Figure 2-5.

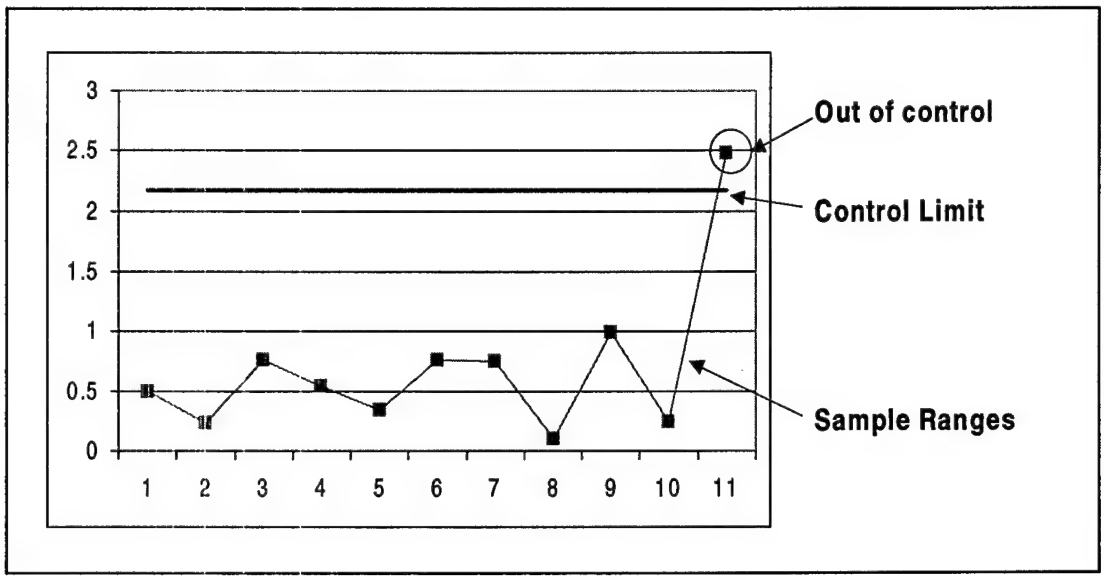

Figure 2-5: Example of an R-chart with an Upper Control Limit

Like the x-bar chart, the R-chart requires that the data is derived from a normal distribution. The range of the sample is calculated by taking the difference between the largest and the smallest value, shown by Equation 2-2.

$$R_i = X_{\max} - X_{\min} \qquad (2\text{-}2)$$

where

R_i = Range of sample i

X_{max} = Largest data point in sample

X_{min} = Smallest data point in sample

The control chart monitors the range of each sample. If the sample range is too great, then the conclusion is that the process is out of control. To calculate the control limits and the centerline, use Equation 2-3 to find the mean of all the ranges.

$$\overline{R} = \frac{R_1 + R_2 + \ldots + R_m}{m} \qquad (2\text{-}3)$$

where

m = number of samples.

$\overline{R}$ = grand mean of all ranges

To find the control limits, we use the assumption of normality to find the standard relationship between the range and the standard deviation. This relationship is often called the relative range, and is defined as $W = R/\sigma$. Using the standard deviation of W, we can estimate the standard deviation of R as $\sigma_R = d_3\sigma$, where d_3 is a function of the sample size n (Montgomery, 1997). Thus, the control limits and the center line are defined by Equation 2-4.

$$\begin{aligned} UCL &= \overline{R} + 3d_3\frac{\overline{R}}{d_2} \\ CL &= \overline{R} \\ LCL &= \overline{R} - 3d_3\frac{\overline{R}}{d_2} \end{aligned} \qquad (2\text{-}4)$$

Simplifying these control limits by letting $D_4 = 1 + 3*(d_3/d_2)$ and $D_3 = 1 - 3*(d_3/d_2)$, we define the control limits by Equation 2-5.

$$\begin{aligned} UCL &= D_4\overline{R} \\ CL &= \overline{R} \\ LCL &= D_3\overline{R} \end{aligned} \qquad (2\text{-}5)$$

Similar to an R-chart is the S-chart. The S-chart also measures the variability of the process, but uses the standard deviation of each sample, versus the range as used by the R-chart. Although the R-chart is easier to calculate than the S-chart, the S-chart has two distinct benefits over the R-chart. The first benefit is that the S-chart is a more accurate representation of process variability than the R-chart. This is especially true for sample sizes greater than ten. The second benefit of an S-chart is that it is better suited for control charts with variable sample sizes. An example of an S-chart is in Figure 2-6:

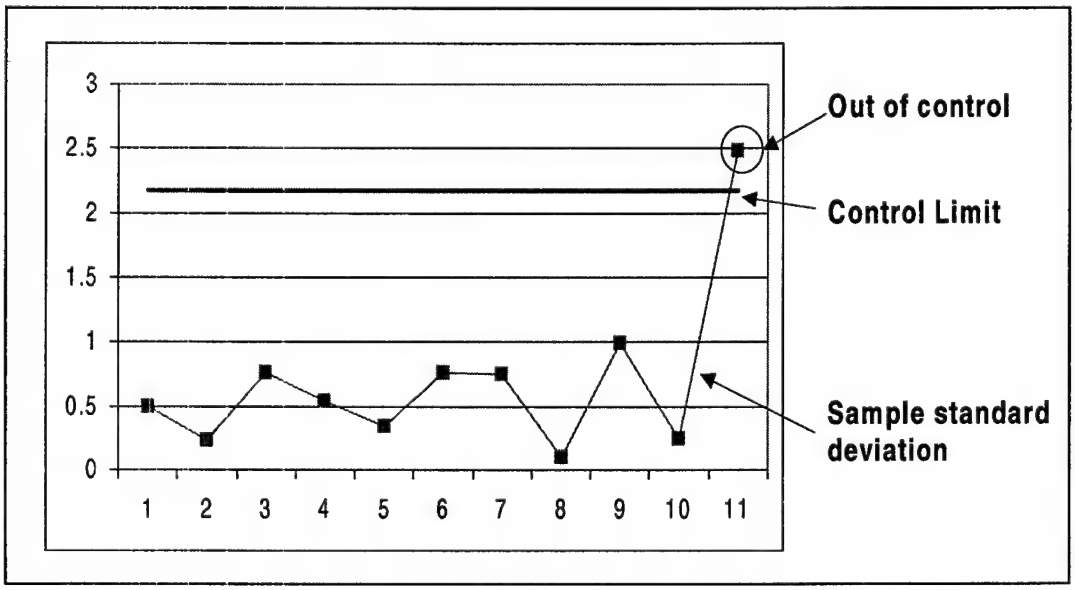

Figure 2-6: Example of an S-Chart with Upper and Lower Control Limits

The sample standard deviation is calculated by taking the square root of the sample variance, which is defined by Equation 2-6.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n-1} \qquad (2\text{-}6)$$

where

x_i = mean of sample I

$\overline{x}$ = grand mean of all samples

n = number of samples

As with the R-chart, the mean standard deviation is simply the average of all sample standard deviations. The control limits are again based upon the relationship between σ and S. For this control chart, $S=c_4\sigma$, where c_4 is a constant based upon the sample size. Also, the standard deviation of S is $\sigma\sqrt{1-c_4^2}$. Therefore, the control limits for an S chart in terms of the sample variance are defined by Equation 2-7.

$$UCL = \overline{S} + 3\frac{\overline{S}}{c_4}\sqrt{1-c_4^2}$$

$$CL = \overline{S} \qquad (2\text{-}7)$$

$$LCL = \overline{S} - 3\frac{\overline{S}}{c_4}\sqrt{1-c_4^2}$$

where

$\overline{S}$ = average of all standard deviations

Again, simplifying the control limits such that $B_3 = 1 - \frac{3}{c_4}\sqrt{1-c_4^2}$ and $B_4 = 1 + \frac{3}{c_4}\sqrt{1-c_4^2}$, the control limits are defined by Equation 2-8.

$$UCL = B_4 \overline{S}$$
$$CL = \overline{S} \qquad (2\text{-}8)$$
$$LCL = B_3 \overline{S}$$

2.4.4. *Variable Charts for Variable Sample Size*

There are few differences between variable charts with constant sample sizes and those with variable sample sizes. The biggest difference is that the control limits for variable sample size control charts are not constant. Since the variance is defined as $\sqrt{\sigma^2/n}$, as the sample size changes, so will the process variance. In order to account for changing sample size, the control limits will differ for each sample. An example of a variable sample size x-bar chart is shown in Figure 2-7:

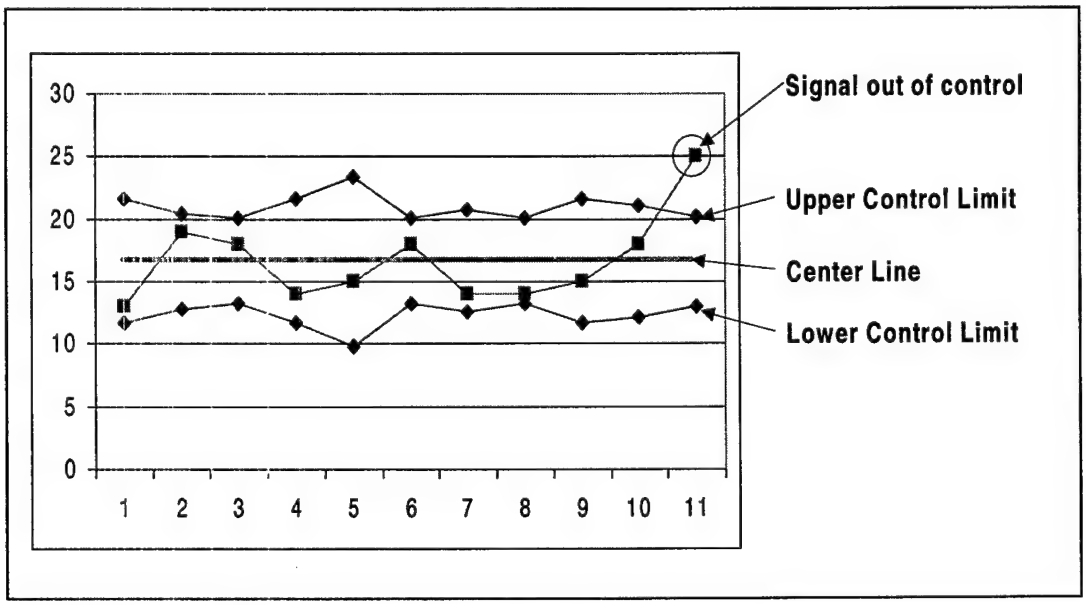

Figure 2-7: Example of an X-bar Chart With Variable Sample Sizes

In addition to the changing control limits, the grand mean is no longer calculated by a simple average of the sample means. Rather, the sample size of each sample must be taken into account. This is done by taking the product of each sample mean and sample size, adding them all together, and dividing it by the sum of all sample sizes. The calculation for the grand mean is defined by Equation 2-9.

$$\mu = \left[\frac{\sum_{i=1}^{m} n_i \bar{x}_i}{\sum_{i=1}^{m} n_i} \right] \qquad (2\text{-}9)$$

where

m = number of samples

n_i = sample size of sample i

$\bar{x}_i$ = mean of sample i

The standard deviation of a variable sample size control chart is defined by Equation 2-10.

$$\bar{S} = \left[\frac{\sum_{i=1}^{m} (n_i - 1) S_i^2}{\sum_{i=1}^{m} n_i - m} \right]^{1/2} \qquad (2\text{-}10)$$

where

m = number of samples

n_i = sample size of i

S_i^2 = variance of sample i

The control limits for a variable sample size chart are based upon the size of each sample. These control limits are defined by Equation 2-11.

$$UCL = \mu + A_3 \overline{S}$$
$$CL = \mu \qquad (2\text{-}11)$$
$$LCL = \mu - A_3 \overline{S}$$

where

μ = grand mean

$\overline{S}$ = standard deviation

A_3 is a relational constant as defined by Equation 2-12.

$$A_3 = \frac{k}{\left[\frac{4(n-1)}{4n-3}\right]\sqrt{n}} \qquad (2\text{-}12)$$

where

n = sample size

k = number of standard deviations between mean and control limit

However, these types of charts do have their limitations. The first limitation is that these charts do not work for autocorrelated data. This is due to the fact that these charts assume that each data point is independent of previous data points. Therefore, time series and correlated data decrease the accuracy of the x-bar chart. The second limitation is that the control chart is built upon data that is assumed to be in control. Since the control limits are based upon the standard deviation of the initial data set, if this data set is not in control, then the control limits are

wider than they should be. Building control limits on data that is not in control in turn decreases the accuracy of the control chart.

2.4.5. The Cumulative-Sum (CUSUM) Charts

Although the x-bar and R charts are effective tools in determining process control, they have two drawbacks. First, these charts are not as effective in detecting smaller shifts, such as a 0.5σ or 1.0σ shift. It takes much longer for an x-bar or an R chart to detect the mean shift, if at all. The second drawback of an x-bar or an R chart is that these charts typically do not take into account previous data points, except in the case of using run rules. A cumulative sum (CUSUM) chart is able to look at historic data to determine if the data trend shows a shift in the data. The CUSUM chart was first developed by Page (1954) and is widely used to monitor the mean of a process. It is better than the standard Shewhart chart in that it is able to detect small deviations from the mean (Reynolds, 1990).

The basic purpose of a CUSUM chart is to track the distance between the actual data point and the grand mean. Then, by keeping a cumulative sum of these distances, we can determine if there is a change in the process mean, as this sum will continue getting larger or smaller. These cumulative sum statistics are called the upper cumulative sum (C_i^+) and the lower cumulative sum (C_i^-). They are defined by Equation 2-13.

$$C_i^+ = \max[0, x_i - (\mu_0 + K) + C_{i-1}^+]$$
$$C_i^- = \max[0, (\mu_0 - K) - x_i + C_{i-1}^-] \qquad \textit{(2-13)}$$

where

x_i = the current data point

μ_o = the grand mean

K = slack value.

However, we do not want the cumulative sums to be too sensitive such that they sum even the smallest deviations from the mean. This is why we include a slack value K. Any deviation that is smaller than the slack value is not added to the cumulative sums. It important to select the right value for K, since a large value of K will allow for large shifts in the mean without detection, whereas a small value of K will increase the frequency of false alarms. Normally, K is selected to be equal to 0.5σ.

Since both the upper and lower cumulative sums are positive numbers, there is only one control limit. This control limit is simply the product of the multiplier (H) and the standard deviation, or $CL = H\sigma$. It is how large the cumulative sum must increase before indicating that the process is out of control. The standard value of H is set to five, thus, most CUSUM control limits are set to five standard deviations. An example of a CUSUM chart is shown in Figure 2-8.

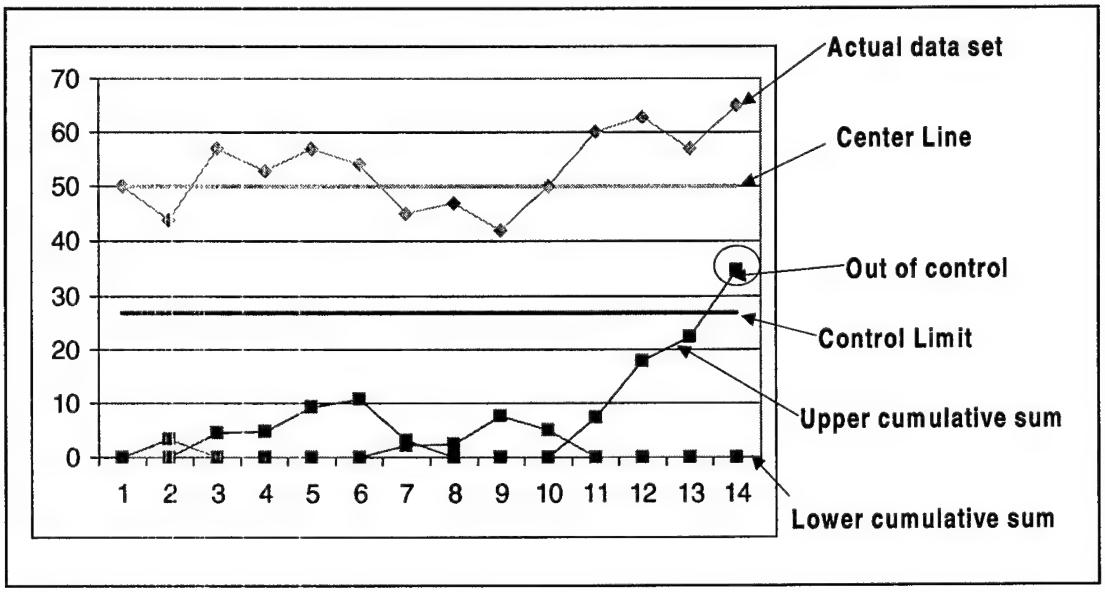

Figure 2-8: Example of a CUSUM Chart

The actual data set has a small mean shift, which occurs at sample eleven. The following points all are above the center line, but not enough to signal on a traditional Shewhart chart.

However, it takes four samples for the CUSUM chart to signal. This is an example of how this type of control chart is an effective method of detecting smaller mean shifts.

2.4.6. The Exponentially Weighted Moving Average (EWMA) Charts

The exponentially weighted moving average chart is similar to the CUSUM chart in that it can detect smaller shifts in the process mean. It also takes into account historic data. This chart tracks the exponentially weighted moving average, which is defined by Equation 2-14.

$$z_i = \lambda x_i + (1-\lambda) z_{i-1} \qquad (2\text{-}14)$$

where

x_i = data point at period i

z_{i-1} = moving average at time i-1

λ = EWMA weight

Lambda (λ) is a weight given to the current data point. Since the value of the moving average is a linear combination of the current data point and the previous moving average, a larger value of λ gives more weight to the current data point. The value of λ can be between zero and one, but is most often chosen between 0.05 and 0.3. The initial value of z (i.e. z_0) is set to the grand mean (μ). Therefore, the current moving average is a linear combination of the current data point and the previous moving average. Since every moving average statistic is a linear combination of the previous moving average, the current moving average is a linear combination of all previous data points. An example of an EWMA chart is shown in Figure 2-9.

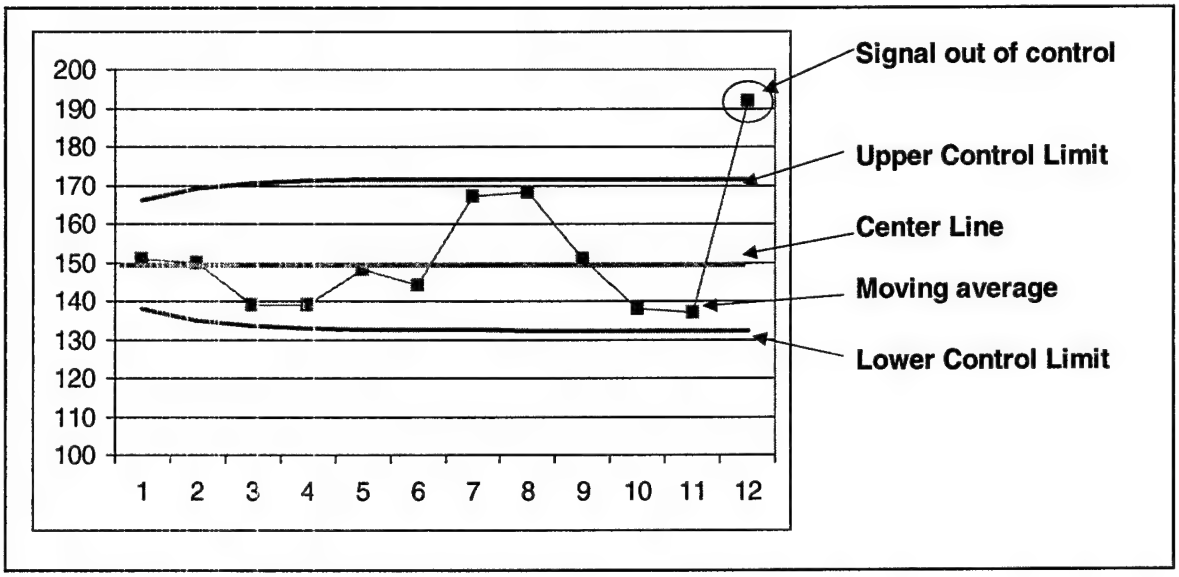

Figure 2-9: An Example of an EWMA Control Chart

The control limits and center line are defined by Equation 2-15.

$$UCL = \mu_o + L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2i}]}$$

$$CL = \mu_0 \qquad (2\text{-}15)$$

$$LCL = \mu_o - L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2i}]}$$

where

μ_o = grand mean

σ = standard deviation

i = sample number

λ = EWMA weight

L = number of standard deviations from the center-line.

2.5. *Autocorrelated Data*

Autocorrelation is a measure of the tendency of neighboring observations from a time-series to vary linearly together rather than independently (Zalewski, 1995). Autocorrelation can lead to false conclusions if it is not accounted for during analysis. One method of determining autocorrelated data is to use the Durbin-Watson test for autocorrelation. Equation 2-16 shows this hypothesis test.

$$H_0 : \rho = 0$$
$$H_a : \rho \neq 0 \qquad (2\text{-}16)$$

where ρ is a measure of the correlation between data points

The larger the value of ρ, the more autocorrelated the data set is. The Durbin-Watson test statistic is the D-statistic, which is calculated by using an ordinary least squares fit and finding the residual e_t, which is defined in Equation 2-17.

$$e_t = Y_t - \hat{Y}_t \qquad (2\text{-}17)$$

where

e_t = the error term or the residual

Y_t = the current data point

$\hat{Y}$ = the estimate of the current data point

The D-statistic is calculated by Equation 2-18.

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad (2\text{-}18)$$

where

n = number of samples

e_t = error term of sample at time t

The smaller the D-statistic is, the more likely the data is autocorrelated. A low value of $e_t - e_{t-1}$ indicates that adjacent error terms tend to be the same magnitude, which is a sign of positive autocorrelation.

Another method of detecting autocorrelation is to look at the correlation between data sets at different lag periods. A high correlation coefficient indicates that there exists a high level of autocorrelation between the data sets at those lag periods. For example, a high correlation coefficient between the data sets one lag apart indicates that the data is highly correlated with the previous time period. By looking at the correlation coefficients of different lag periods, we hope to determine the number of lags needed such that there is no autocorrelation.

2.5.1. Control Charts for Autocorrelated Data

The presence of autocorrelated data affects the reliability of the traditional control chart. The most common result is the increase in the false alarm rate. This means that the control chart will signal that the process is out of control much more often. Most of these signals will not have assignable causes; rather, they are induced by the autocorrelative structure of the data (Mastrangelo and Montgomery, 1991).

The general solution to the problem of autocorrelation is to fit an appropriate time series model and apply a control chart to the residuals. One of the most common time series models to

use with an autocorrelated data set is a first order autoregressive model, or an AR(1) model. It assumes that each data point is defined by Equation 2-19.

$$X_t = \xi + \phi X_{t-1} + \varepsilon_t \qquad (2\text{-}19)$$

To find the values of ξ and ϕ, use a least squares linear fit on the data set. The residuals (ε_t) should now be independent and normally distributed.

Another method is to use a moving centerline EWMA chart. This uses the value of the moving average Z to determine control limits for the actual data. To determine the control limits for X_t, calculate the value of Z_{t-1} using equation 2-14. This value of Z_{t-1} becomes the center line for X_t. Thus, the control limits are calculated using Equation 2-20

$$\begin{aligned} UCL_t &= z_{t-1} + 3\sigma \\ CL_t &= z_{t-1} \\ LCL_t &= z_{t-1} + 3\sigma \end{aligned} \qquad (2\text{-}20)$$

Thus the center line and control limits change according to the previous moving average. An example of this type of control chart is shown in Figure 2-10

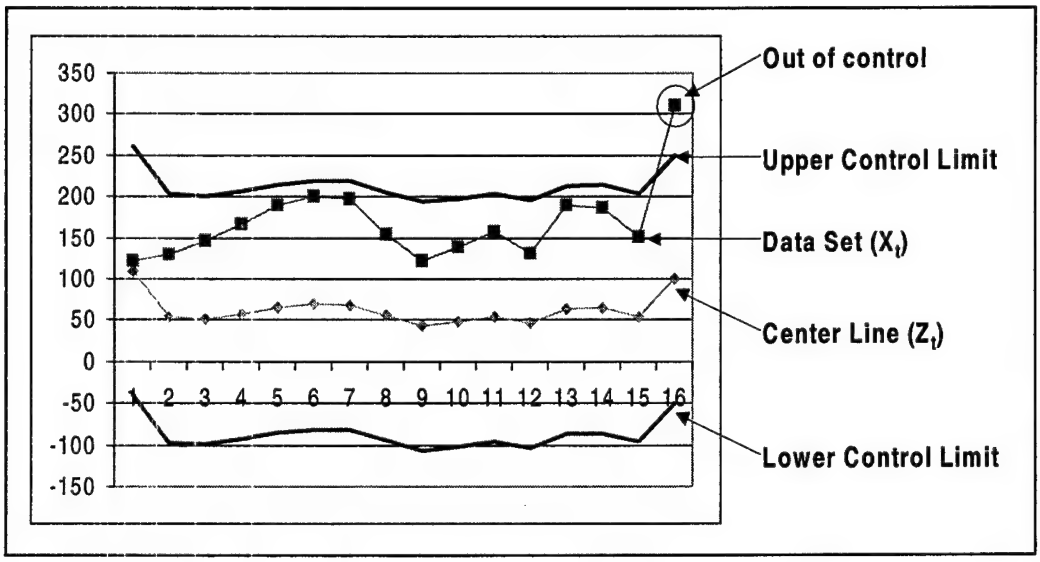

Figure 2-10: Example of a Moving Centerline EWMA Chart

However, these charts are not effective when used to classify pilot mental workload. These charts do not signal during periods of high workloads. Instead, they signal when there appears to be a change in the workload level. Figure 2-11 shows actual heart beat interval data. Notice that the chart does not signal very often. However, those signals indicate changes to the workload level.

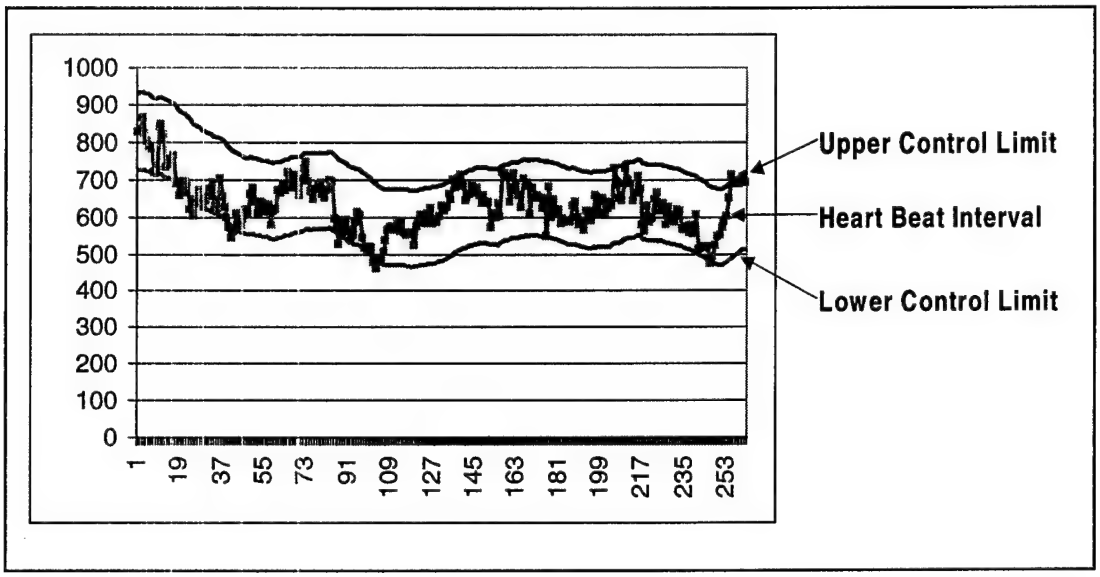

Figure 2-11: Moving Centerline EWMA Chart Using Heart Beat Interval for Pilot 1, Day 1

A simple way to use an EWMA chart with autocorrelated data is to let $\lambda=1 - \theta$, where θ is found using a first-order integrated moving average model, shown in Equation 2-21.

$$X_t = x_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1} \qquad (2\text{-}21)$$

This is only an approximation, and assumes that the first-order integrated moving average is the underlying time series function. It also assumes that the data is positively autocorrelated, and that the process mean does not drift too quickly.

3. Traditional Control Charts

3.1. Overview

This chapter looks at the use of different control charts as classifiers of pilot mental workload. The first control chart we use to identify pilot mental workload is the standard x-bar chart. We use this chart because it is relatively simple to build and to interpret (Montgomery, 1997). We find that the most challenging part of building the control chart is the conversion of the data set. This involves collecting the raw data and grouping it into samples. There is also a variable that has a high level of autocorrelation. Since standard Shewhart charts assume independent data, we address this problem by trying several other control charts. Once we build these control charts, we also vary the control limits to maximize the classification accuracy. Finally, we examine whether a combination of the seven control charts may improve the classification accuracy.

3.2. Data Collection and Conversion

The data sets used for this research consist of 151 different psychophysiological features. These features were monitored and recorded for one pilot during a forty-four-minute flight. The flight was split into twenty-two, two-minute segments, each of which was classified as "high" or "low" workload. Of the twenty-two flight segments, 58% of them were classified by the pilot as low workload, and 42% of the segments were classified as high workload. Figure 3-1 shows the actual workload level across all twenty-two segments.

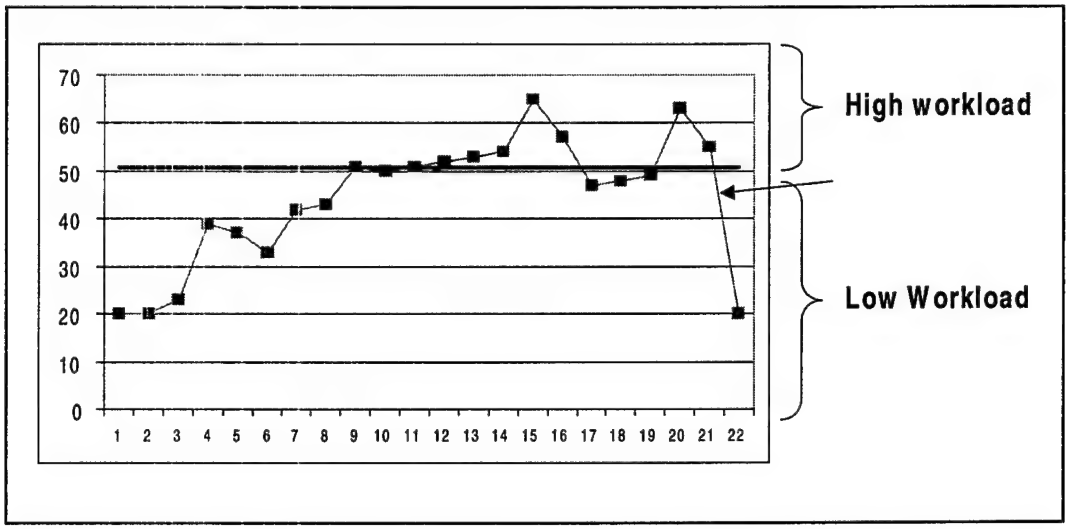

Figure 3-1: Pilot Mental Workload Level Across All Twenty-Two Segments

Data were collected for ten different pilots on two different days, with each pilot flying the same route on each day. This research effort uses the flight data for pilot 1 and pilot 4 across two separate days, for a total of four different data sets. Each data set contains the brain activity, cardiac, respiratory, and ocular measurements that were recorded during the flight. For cardiac measures, the data set contains the time between heartbeats. For respiratory measurements, the data set contains the time between breaths, and the minimum and maximum amplitude of the breath. For ocular measurements, the data set has time between blinks, the amplitude of the blink, and the duration of the blink.

3.2.1. Developing Synchronized Data Set

One of the classification schemes we are using is a combination of the different psychophysiological features. However, heartbeats, eye blinks, and breaths do not occur at the same time, and therefore there can be no comparison between individual events. Rather, the comparison between features must be within a given time period (i.e. what are these features doing in the first minute of flight?). We must sample the data such that the duration of each

sample is constant. However, constant time duration means that the size of each sample will be different.

The program that compiles the raw data, generates the synchronized data set, and puts the data into a spreadsheet is a Microsoft Excel macro. The macro reads in the data for each psychophysiological feature and separates the data into samples, each with the same time interval. One of the considerations in building this data set is how to determine the length of the time interval. It has to be long enough so that on the average, there are more than two data points per sample. This requirement provides a measure of the data variability, which is critical in building the control limits. Of the average arrival times of heartbeats, breaths, and blinks, we find that the most infrequent feature to occur are eye blinks, which have an average inter-arrival time of roughly 4.7 seconds. To ensure that a majority of the samples have at least two eye blinks, the minimum time interval must be at least 9.4 seconds. However, a larger time interval decreases the response time of the classifier, so we want to keep this time interval as short as possible. For this research effort, we are using a time interval of ten seconds. Although taking samples at such a short time interval decreases the accuracy of the estimate of the process variation, a longer time interval delays the response time of the classifier. The decreased response time affects its utility when used with the automated system.

For each sample, the macro calculates the sample size (n), the sample mean ($\overline{X}$), the sample standard deviation (S), and the appropriate value for A_3. The value for A_3 is dependent upon sample size, and is used to calculate the control limits.

3.3. Classification Accuracies

For each classifier, we generate three classification statistics to compare the utility of one classifier against other classifiers, both past and present. The three statistics are misidentification rate, false alarm rate, and overall classification rate. A misidentification rate is defined as the proportion of high workload samples that are classified as low workload. This is calculated by

taking the number of misidentified high workload samples and dividing by the total number of high workload samples. The false alarm rate is defined as the proportion of low workload samples that are classified as high workload. This is calculated by taking the number of misidentified low workload samples and dividing it by the total number of low workload samples.

An easier way to look at these two statistics is to think of them as type I and type II errors. Since a control chart is a series of hypothesis tests, we define the null hypothesis test as $\mu=\mu_o$, or in simpler terms "the workload is low." By this definition, a type I error would be defined as "classifying workload as high when the workload is actually low." This is the same definition as a false alarm rate. A type II error would be defined as "classifying workload as low when the workload is actually high." This is how we define the misidentification rate.

The overall classification accuracy is the percentage of the samples that are correctly identified. Using the same hypothesis test defined in the previous paragraph, this is defined as the combination of "classifying workload as high when the workload is actually high," and "classifying workload as low when the workload is actually low." To calculate this accuracy, take the sum of the number of correctly identified low and high workload samples and divide the sum by the total number of samples. This is a very important metric, as this allows us to compare our SPC classifiers against multivariate discriminant classifiers. The classifiers that East (2000) built had classification accuracies between 74% and 97% when used to classify workload for same pilot on the same day. The classification accuracies varied depending upon the type of classifier used and the data set. The results from her research are shown in Table 3-1

Table 3-1: Summary of Classification Accuracies Using Multivariate Discriminant Classifiers

Data Set	Initial	SAS	SNR	Factor Analysis	Factor Analysis
Pilot 1, Day 1	78.50% (151 vars)	82.51% (34 vars)	81.79% (14 vars)	77.79% (10 vars)	74.07% (6 vars)
Pilot 1, Day 2	75.52% (151 vars)	78.51% (71 vars)	77.16% (17 vars)	75.15% (13 vars)	
Pilot 4, Day 1	96.72% (146 vars)	97.46% (79 vars)	91.78% (5 vars)	90.41% (3 vars)	
Pilot 4, Day 2	86.91% (151 vars)	90.50% (62 vars)	85.76% (5 vars)	85.92% (3 vars)	

SPC classifiers must have this level of classification accuracy if they are to be considered comparable classifiers to multivariate discriminant classifiers. Also, each classifier has a minimum classification accuracy that can be obtained by simply classifying all segments as low workload. This results in a false alarm rate of 0% and a misidentification rate of 100%. Since 58% of the segments are classified as low workload, this classifier will result in a classification accuracy of 58%. We will call this the naive classifier, as it requires no effort or analysis to achieve this level of classification accuracy. The classification accuracy of the naive classifier is significant for two reasons. First, several of the control charts do not produce a signal. This chart can be considered a naïve classifier, because it classifies every segment as low workload. The second significance is that several of these control charts will actually produce a classification accuracy of less than 58%. Since a naïve classifier would produce better results, a control chart that has a classification accuracy of less than 58% is considered to be insignificant and is thrown out.

3.4. Initial data analysis

3.4.1. Autocorrelation of the Data Set

Before starting the analysis of the data, we checked to see if any of the psychophysiological features had autocorrelated data. We expect to see some degree of

autocorrelation among the inter-arrival times of the different factors. These include the heart beat interval, the breath interval, and the blink interval. To determine if any of the factors were indeed autocorrelated, we looked at the correlation of the data at lag 1 of all seven features for both pilots on both days. The results are shown in Table 3-2.

Table 3-2: Correlation of Seven Features at Lag 1 for Both Pilots, Both Days

Data Set	Heart Beat Int	Breath Int	Breath Min	Breath Max	Blink Int	Blink Amp	Blink Dur
Pilot 1, Day 1	.953	.276	.207	.351	.236	.028	.039
Pilot 1, Day 2	.968	.117	.235	.274	.195	.383	-.002
Pilot 4, Day 1	.984	.204	.138	.325	.149	.338	.042
Pilot 4, Day 2	.952	.230	0	-.011	.093	.327	.040

The only factor that shows significant signs of autocorrelation is heart beat interval. This is also the only psychophysiological feature that was highly correlated across all four data sets. Although it is not surprising that the heart beat interval was highly autocorrelated, we did not expect the low levels of correlation with the breath interval and blink interval. This analysis also shows that the breath maximum amplitude and blink amplitude may be autocorrelated as well. However, the level of correlation at lag 1 is not significant for these features. It was also found that these features are not autocorrelated across all four data sets. For this research, we assume that these features are not autocorrelated. We decided that the best way to deal with the autocorrelated data is to build an EWMA chart, since this chart accounts for data trends. We will use standard x-bar charts on the other six non-correlated features.

3.4.2. Finding the Low Data Set

One of the assumptions of a control chart is that the mean and control limits are defined by data that is in control. If we use control charts to tell the difference between low and high workload segments, then low workload segments are defined as in-control and high workload

segments are defined as out-of-control. Since the data set contains segments that are classified as both high and low workload, building the control chart using the entire data set would produce inaccurate control limits. The solution is to establish a subset of the data that consists of only low workload segments. We call this subset the baseline of the data set. From this baseline, we calculate estimates of the mean and standard deviation, which are important when building different control charts. This estimate is specific to that feature for that pilot on that day. This is important because when we apply different classifiers to this data set, we change different variables, but the estimates of the mean and standard deviation do not change.

However, the transitions between low and high workload segments are not instantaneous. During some of the segments classified as low workload, the pilot may start to show signs of high workload. This is because the pilot is either gearing up for these anticipated mental demands or the body is still recovering from a previous one. There should also be enough data points to generate a statistically significant baseline data set. This baseline should have at least thirty samples to produce an accurate estimate of the mean and standard deviation (Montgomery, 1997). Finally, the baseline data set should be collected at the beginning of the flight. Since the baseline determines the appropriate levels for the control limits, the control charts do not start monitoring the psychophysiological features until after the baseline is complete.

The predetermined flight path showed that the first eight segments of flight (first sixteen minutes) are all considered low workload. However, since the ninth segment is classified as high workload, the eighth segment may start to show signs of building up to the high workload segments. Since each sample is ten seconds long, at least three flight segments are needed to generate thirty samples. The baseline data set needed to be between the first three and first seven segments of the flight as the baseline. To determine which segments to use as the baseline, we took several psychophysiological features from the Pilot 1, Day 1 data set, and built an x-bar chart. For each x-bar chart, we change the control limits based upon the estimates of the mean and standard deviation generated from the baseline samples. We chose the first four

segments (first eight minutes of flight) as the baseline, because it had the highest classification accuracies across the different features.

3.4.3. Classifying Pilot Mental Workload Across Different Pilots and Days

When we take a classifier and test it against other pilots and days, we take the estimates of the mean and standard deviation for that data set and keep it constant for all classifiers. Since the estimates of the mean and standard deviation do not change, what does change are certain control chart variables. These variables are used to determine the distance between the control limits and the center line. This distance is in terms of the standard deviation of the process, and these variables determine how many standard deviations there are between the center line and each control limit. As the multiplier increases, the distance between the center line and the control limits also increases. To use a classifier to classify mental workload for other pilots or days, we use the multiplier of the classifier with the estimates of the mean and standard deviation of the data set to classify.

3.5. X-bar Charts for Non-correlated Data

For the each of the six non-autocorrelated features, we will build a standard x-bar chart, using the data set we separated into ten second intervals. Since the data set had variable sample sizes, we calculate the mean and standard deviation using Equations 2-9 and 2-10 on the first eight minutes of flight. Initially, we built the x-bar charts using the typical three-sigma control limits. We found that these x-bar charts produced low false alarm rates, but generated high misidentification rates, which results in marginal classification accuracies. For Pilot 1, Day 1, the x-bar chart that used inter-blink interval as the sole psychophysiological feature yielded the highest classification accuracy. However, this classification accuracy was only 63.3%, which is lower than the 74% classification accuracy produced by the worst multivariate classifier (East,

2000). The classification accuracies using single psychophysiological features with x-bar charts are shown in Table 3-3.

Table 3-3: False Alarm Rates, Misidentification Rates, and Classification Accuracies for Pilot 1, Day 1 Based Upon Single Feature X-bar Charts with Three-Sigma Control Limits

Feature	False Alarm Rate	Misidentification of High Workload Rate	Overall Classification Accuracy
Breath Interval	3.27%	7.21%	59.09%
Breath Min Amp	13.73%	9.91%	54.17%
Breath Max Amp	3.92%	4.50%	57.58%
Blink Interval	22.88%	25.23%	55.30%
Blink Amplitude	5.23%	9.01%	58.71%
Blink Duration	5.23%	7.21%	57.95%

None of these six control charts produces a classification accuracy over 60%. We build the same six control charts using the other three data sets, and found similar results. The results from all four data sets are shown in Appendix A.

Since none of these six control charts produced adequate classification accuracies, we rebuilt each x-bar chart but allowed the value of k to change. Changing the value of k changes the control limits, which impacts the classification accuracy. For each control chart, we changed the value of k to optimize the classification accuracy. The classification accuracies using single psychophysiological features with variable control limit x-bar charts are shown in Table 3-4.

Table 3-4: False Alarm Rates, Misidentification Rates, and Classification Accuracies for Pilot 1, Day 1 Based Upon Single Feature X-bar Charts with Variable-Sigma Control Limits

Feature	Value of k	False Alarm Rate	Misidentification of High Workload Rate	Overall Classification Accuracy
Breath Interval	1.6	3.27%	84.68%	62.50%
Breath Min Amp	3.0	13.73%	9.91%	54.17%
Breath Max Amp	3.0	3.92%	4.50%	57.58%
Blink Interval	5.0	16.99%	74.77%	58.71%
Blink Amplitude	9.0	30.72%	18.92%	74.20%
Blink Duration	2.5	38.56%	9.01%	73.86%

We found that the classification accuracies improve for four of the six of the x-bar charts. The most significant improvements are found using blink amplitude and blink duration x-bar charts. The classification accuracy of these two charts improved to over 73%, which is very comparable to previous classifiers. The results from the other three data sets are in Appendix B

Again, we build the same control charts for the other three data sets. For each data set, we maximize the classification accuracy for each psychophysiological feature by changing the value of k. We find that the classification accuracies for the other three data sets are very similar to ones generated from pilot 1, day 1. The only x-bar chart that that produces decent levels of classification accuracies across all four data sets was one used with blink duration. This x-bar chart has classification accuracies that exceeded 68% across all four data sets. The classification accuracy using blink duration with an x-bar chart is shown in Table 3-5.

Table 3-5: Classification Accuracies of X-bar Charts Using Blink Duration as the Sole Feature

Data Set	Value of k	False Alarm Rate	Misidentification of High Workload Rate	Classification Accuracy
Pilot 1, Day 1	2.5	38.56%	9.01%	73.86%
Pilot 1, Day 2	1.0	37.75%	22.73%	68.58%
Pilot 4, Day 1	1.3	16.77%	35.19%	75.67%
Pilot 4, Day 2	1.3	17.42%	30.28%	77.27%

An x-bar chart using blink duration may be a good classifier of pilot mental workload. However, ocular measurements are good indicators of visual information processing. Although some instances of high mental workload is a result of high visual demands, there are instances where high mental workload is a result of other types of stressors. It may be during these instances where this control chart fails to accurately classify pilot mental workload.

3.6. EWMA Control Charts for Autocorrelated Data

X- bar charts do not work with autocorrelated data, so we use a different type of control chart to monitor heart beat interval data. One control chart that works well with autocorrelated

data is an EWMA chart (Mastrangelo and Montgomery, 1991). To build this chart, we converted the mean of each sample to a Z statistic, which is defined by Equation 2-16. To build the control limits for the EWMA chart, we choose typical values for L. To provide an estimate of λ, we use Equation 2-21 to find an estimate of θ. We estimate λ to be $1-\theta$. The results of using an EWMA chart across all four data sets are shown in Table 3-6.

Table 3-6: False Alarm Rate, Misidentification Rate, and Classification Accuracy for a Standard Three Sigma EWMA Chart for All Four Data Sets

Data Set	False Alarm Rate	Misidentification of High Workload Rate	Overall Classification Accuracy
Pilot 1, Day 1	45.75%	32.43%	59.85%
Pilot 1, Day 2	23.84%	60.91%	60.54%
Pilot 4, Day 1	27.10%	12.96%	78.71%
Pilot 4, Day 2	37.42%	0.92%	77.65%

Although the EWMA chart has marginal results for both pilot 1 data sets, the same chart for both pilot 4 data sets attain classification accuracies over 77%. We also maximize the classification accuracy by changing the values for L. Again, we find that the classification accuracy for pilot 1 was still low, but the classification accuracy for pilot 4 was over 81% on both days. The results of this analysis are shown in Table 3-7.

Table 3-7: False Alarm Rate, Misidentification Rate, and Classification Accuracy for a Variable-Sigma EWMA Chart for All Four Data Sets

Data Set	Value of L	Value of λ	False Alarm Rate	Misidentification of High Workload Rate	Overall Classification Accuracy
Pilot 1, Day 1	9.1	0.0271	45.10%	33.33%	59.85%
Pilot 1, Day 2	5.5	0.0202	36.42%	43.64%	60.54%
Pilot 4, Day 1	7.1	0.0501	16.13%	21.20%	81.75%
Pilot 4, Day 2	7.1	0.0662	21.94%	8.26%	83.71%

Even though the EWMA chart using Pilot 4 data produces good classification accuracies, there is no evidence that this control chart is an accurate method of classifying mental workload for all pilots. We know that using an EWMA chart with cardiac features using pilot 1 data would only result in a classification accuracy of 60.54% at best. Therefore, EWMA charts do not seem to work across pilots.

3.7. Monitoring Multiple Control Charts

For each of the four data sets, we use the EWMA chart for heart rate with the other six x-bar charts to see if a combination of these seven control charts could improve the accuracy of the classifier. The control charts with variable-sigma control limits has better classification accuracies than standard three-sigma control charts, so we use these control charts only. We look at 120 possible combinations of the seven control charts to determine if there is a combination that provided significant improvement to the current classifier. Each combination included anywhere between two and seven control charts. We assume that for there to be a signal, all the control charts must signal during the same segment. We found that of the 120 different combinations of control charts, ninety-three of them did not signal at all. Therefore, these classifiers have the same classification accuracy as the generic classifier (0% false alarm rate, 100% misidentification rate). Of the remaining twenty-seven, none of the combined control charts has a classification accuracy greater than 63.3%. The only promising classifier was a combination of all seven different control charts.

This combination of control charts can be considered as series of binomial random variables (i.e. of the seven control charts, x control charts signaled). For each sample, we count the number of control charts that signals a high workload. High workload is signaled if a certain number of control charts signal in the same sample period. We vary the number of control chart needed to signal for a high workload to optimize the classification accuracy. The result of this analysis is in Table 3-8.

Table 3-8: False Alarm Rates, Misidentification Rates, and Classification Accuracies for Pilot 1, Day 1 Based Upon Number of Charts Signaling

Number of Charts to Produce a Signal	False Alarm Rate	Misidentification of High Workload Rate	Classification Accuracy
1 of 7 signals	37.25%	78.38%	45.45%
2 of 7 signals	18.95%	79.28%	55.68%
3 of 7 signals	18.30%	79.28%	56.06%
4 of 7 signals	18.30%	97.30%	48.48%
5 of 7 signals	5.88%	100%	54.55%
6 of 7 signals	0.65%	100%	57.58%
7 of 7 signals	0%	100%	57.95%

We found that the classification accuracies decrease when using multiple control charts. Similar results are found using the other three data sets. Using a combination of x-bar and EWMA charts proved to be less accurate than a control chart for individual features.

3.8. Cumulative Sum (CUSUM) Charts

Since x-bar charts did not provide the desired classification accuracy, we use a different type of control chart. A cumulative sum (CUSUM) chart might work better, since it has some advantages over x-bar charts. The first advantage is that it is able to detect smaller mean shifts. Also, the time between the actual mean shift and when the control chart signals is reduced. This means that a CUSUM chart should be able to detect high workload periods quicker, thus reducing the misidentification rate. The second advantage is that a CUSUM chart looks at historic data, whereas an x-bar chart assumes independent samples. The classification of high mental workload may be better classified using data trends or data runs, rather than independent data samples.

Although there has been work done using variable-size samples for Shewhart charts, there is little research in the realm of a variable-size CUSUM chart. Most of the research in this area deals with the use of varying the sample size (Annadei et al, 1995) or sample interval (Reynolds et al, 1990) to improve the accuracy and performance of the CUSUM chart. However, the assumption with these methods is that it allows the analyst to adjust the sample size based

upon past results. It does not address the issue of data that has naturally varying sample size, as in the case with our data set. Therefore, we have to try a different method to resolve the problem of multiple sample size.

The method we chose to calculate the mean and standard deviation is to use the same method for calculating a mean and standard deviation as a variable size Shewhart chart. Thus, for this CUSUM chart, we calculate the mean using Equation 2-9 and the standard deviation with Equation 2-10. We also use the same first four segments as the data set as the baseline.

3.8.1. Traditional CUSUM Charts

Using the same synchronized data set used with the Shewhart charts, we build seven CUSUM charts for each data set to determine the effectiveness of the CUSUM chart. We build these control charts using the values of $k = 0.5\sigma$ and $H = 5$, which are typical for CUSUM charts. Using the standard value for H, we found that the control charts did not have very high classification accuracies. Therefore, we optimize the classification accuracy for each individual psychophysiological feature by adjusting the value of H. The results of this analysis are shown in Table 3-9.

Table 3-9: False Alarm Rates, Misidentification Rates, and Classification Accuracies for Pilot 1, Day 1 Using CUSUM Charts

Feature	Value of H	False Alarm Rate	Misidentification of High Workload Rate	Classification Accuracy
Heart Beat Interval	53	36.60%	0%	78.79%
Breath Interval	0.3	3.92%	83.78%	62.50%
Breath Min Amp	40	0%	100%	58.02%
Breath Max Amp	70	0%	100%	58.02%
Blink Interval	3.1	12.42%	75.68%	60.98%
Blink Amplitude	350	0%	100%	58.02%
Blink Duration	50	0%	100%	58.02%

By building the control charts in this method, we found that there was a big improvement over traditional Shewhart charts in terms of classification accuracy. In fact, the use of heart beat interval alone had better classification accuracies than previous classifiers. We found this to be true across all four data sets. These results are shown in Table 3-10.

Table 3-10: Results of Using a CUSUM Chart with Heart Beat Interval Across Both Pilots, Both Days

Data Set	Value of H	False Alarm Rate	Misidentification of High Workload Rate	Overall Classification Accuracy
Pilot 1, Day 1	53	36.60%	0%	78.79%
Pilot 1, Day 2	53	3.95%	24.55%	87.40%
Pilot 4, Day 1	48	38.06%	0%	77.57%
Pilot 4, Day 2	17	38.06%	0%	77.65%

3.8.2. Combination of CUSUM Charts

As with the Shewhart charts, we looked to see if using multiple CUSUM charts could improve the classification accuracy. The results are shown in Table 3-11.

Table 3-11: False Alarm Rates, Misidentification Rates, and Classification Accuracy of Pilot 1, Day 1 Based Upon Number of Charts Signaling

Number of Charts to Produce a Signal	False Alarm Rate	Misidentification of High Workload Rate	Classification Accuracy
1 of 7	60.78%	24.32%	54.55%
2 of 7	8.50%	90.99%	56.82%
3 of 7	0%	100%	57.95%
4 of 7	0%	100%	57.95%
5 of 7	0%	100%	57.95%
6 of 7	0%	100%	57.95%
7 of 7	0%	100%	57.95%

We found that combining the CUSUM charts actually diminishes the classification accuracy. We found similar results when combining the seven CUSUM charts for the other four data sets. It

appears that the combination of control charts increases the error rates, rather than the classification accuracies. Therefore, it may be more beneficial to look individual psychophysiological features rather than a combination of them.

3.9. Testing the Classifiers Across Pilots and Days

Although we found several classifiers that had high classification accuracies for the same day and same pilot, the next step is to determine if these classifiers can retain these high classification accuracies across the different data sets. To apply the control charts across the different data sets, we take the values of K or H (difference between the center line and the control limits) and use them to rebuild the control limits for the other data sets. However, we still use the value of the mean and standard deviation specific to that data set. These numbers are calculated using the first four segments of flight, which we established as the baseline.

3.9.1. CUSUM Chart Using Heart Beat Interval

The first control chart we test was the CUSUM chart using heart beat interval. The value of H was specific to the classifying data set, but the mean and standard deviation were specific to the data set being classified. The result of this analysis is in Table 3-12.

Table 3-12: Classification Accuracy Using Heart Beat Interval as a Classifiers Across Different Pilots and Different Days

Data Set	Classification Accuracy Against Pilot 1, Day 1	Classification Accuracy Against Pilot 1, Day 2	Classification Accuracy Against Pilot 4, Day 1	Classification Accuracy Against Pilot 4, Day 2
Pilot 1, Day 1	78.8%	85.6%	73.1%	75.4%
Pilot 1, Day 2	77.8%	86.4%	73.9%	74.6%
Pilot 4, Day 1	65.2%	75.4%	76.1%	75.4%
Pilot 4, Day 2	58.0%	76.1%	67.4%	76.9%

We found that the classification accuracy decreases when applied to the other data sets. However, the Pilot 1 classifiers maintained a high classification accuracy across all four data sets. In both cases, the classification accuracy remained above 73%. This is a big improvement over past classifiers in that the classification accuracy does not decrease significantly when applied to different pilots and days.

3.9.2. X-bar Chart Using Blink Duration

The second control chart we found had decent classification accuracies across the different data sets was an x-bar chart using blink duration as the primary classifier. To apply the classifier, we changed the value of K, but kept the value of the mean and standard deviation specific to the data sets. The result of the analysis is in Table 3-13.

Table 3-13: Classification Accuracy of Using Blink Duration as a Classifier Across Different Pilots and Different Days

Data Set	Classification Accuracy Against Pilot 1, Day 1	Classification Accuracy Against Pilot 1, Day 2	Classification Accuracy Against Pilot 4, Day 1	Classification Accuracy Against Pilot 4, Day 2
Pilot 1, Day 1	73.86%	65.90%	65.80%	72.35%
Pilot 1, Day 2	68.56%	68.58%	75.29%	75.76%
Pilot 4, Day 1	70.83%	68.58%	75.63%	77.27%
Pilot 4, Day 2	70.83%	68.58%	75.63%	77.27%

We found that the classification accuracies did not significantly decrease when applied to other data sets. We found that the classifier for Pilot 4 seemed to work the best across the different data sets. These classifiers did not produce the same classification accuracies as previous classifiers when used to classify pilot mental workload for the same pilot on the same day. However, they maintain the classification accuracy when used to classify pilot mental workload for different pilots on different days.

3.10. Sumrnary

We found that using traditional control charts for variables is not an effective classifier with regard to pilot workload. In only a few cases did these control charts work. However, what we did find is that the control charts classifiers produces high classification accuracies maintained a high level of classification accuracy across the different data sets. We found that what a control chart lacks in same day, same pilot accuracies it makes up for by being robust enough to classify across different pilots and different days.

4. Other Types of Control Charts

4.1. Overview

Although several of the traditional control charts are acceptable classifiers of pilot mental workload, we want to make several modifications to the traditional control charts. One of the modifications we make is to modify the traditional cumulative sum (CUSUM) chart to include a ceiling. This ceiling limits the value of the cumulative sum statistics. This should make the CUSUM chart more sensitive to a change from high mental workload to low mental workload. Another change is to rebuild the data set so that the samples are of a fixed size, as opposed to samples that are of a fixed time interval. Since we found that control charts using singular psychophysiological features are more accurate, there is no need to keep the constant time interval.

4.2. Cumulative Sum (CUSUM) Charts With a Ceiling

A CUSUM chart has several advantages over Shewhart charts. First, it is an effective method of detecting small changes in the process mean. It is very possible that the psychological differences between high and low workload segments are too small to be detected by an x-bar chart. A CUSUM chart is very effective in detecting shifts between one and two standard deviations (Reynolds et al, 1990). Also, a CUSUM chart might be very effective with dealing with the ramp up periods before and after high mental workload segments. Since the CUSUM is able to detect data trends, the ability to look at previous data points will help classify pilot workload.

The misidentification rate of the best classifiers is much lower than the false alarm rate. Improvement to the classification accuracy requires minimizing the occurrence of this type I error. One of the reasons why the false alarm rate is larger than the misidentification rate is the CUSUM chart is unable to track the low workload periods between two separate high workload periods. During these high workload periods, the values of the cumulative sum statistics increases so

rapidly that it takes too long to come back down below the control limit and indicate the pilot has returned to a state of low mental workload. Often, by the time that the cumulative sum statistic decreases enough to cross the control limit again, the pilot begins another period of high mental workload. The result is a high occurrence of false alarms. A limit on the cumulative sum statistic would decrease the response time to a period of low workload. This decreases the occurrence of false alarms and increases the accuracy of the classifier.

The biggest question with this type of control chart is to determine where the ceiling should be in relationship to the control limit. Since there has been no research in this area, we will try different distances for the ceiling height (c). An example of the CUSUM chart with a ceiling is shown in Figure 4-1.

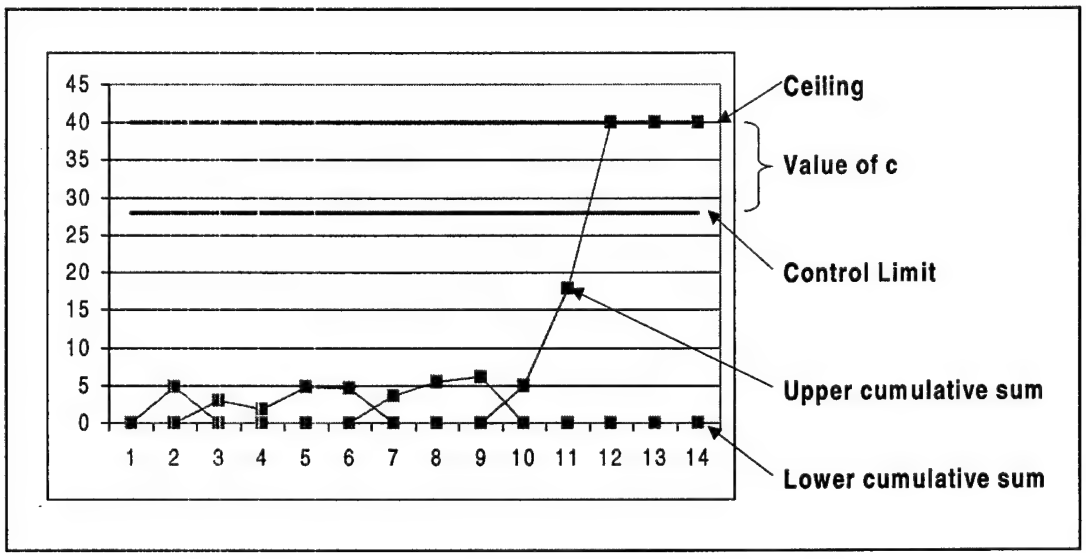

Figure 4-1: Example of a CUSUM Chart with a Ceiling

We initially choose a value of $c=10\sigma$ for the ceiling height for all data sets. Since heart beat interval was the only psychophysiological feature to have significant classification accuracies, we only built modified CUSUM charts with this feature. The classification accuracies of the modified

CUSUM chart are compared to the results of a traditional CUSUM chart. The result of this are in Table 4-1.

Table 4-1: Comparison of Traditional and Modified CUSUM Charts Using Heart Beat Interval Based Upon Classification Accuracy

Data Set	Traditional CUSUM Classification Accuracy	Modified CUSUM Classification Accuracy
Pilot 1, Day 1	78.79%	78.78%
Pilot 1, Day 2	87.40%	87.40%
Pilot 4, Day 1	77.57%	86.69%
Pilot 4, Day 2	77.65%	77.65%

In most cases the modified CUSUM chart did not increase the classification accuracy of the control chart. One possible conclusion is that the classification accuracy is sensitive to the value of c. To test this theory, we rebuild each CUSUM chart to maximize the classification accuracy, subject to both the value of H and c. We find that only in one of the four cases did the classification accuracy improve, and this improvement was very minimal. Therefore, the classification accuracy does not seem to be very sensitive to the ceiling height.

The modified CUSUM charts are not as effective as we originally hoped it might be. This new modification to the CUSUM chart may need to be reworked before it can be a useful tool to classify pilot mental workload. Therefore, until we can make additional improvements to the modified CUSUM chart, using traditional CUSUM charts seems best.

4.3. CUSUM Charts with Fixed Sample Sizes

We found that a CUSUM chart using heart beat interval is able to classify pilot mental workload. Therefore, we want to explore different variations on this classifier to improve the classification accuracy. The original data set uses a constant time interval but varies the sample size. However, there is no reason to keep this time interval if heart beat interval is not used in conjunction with any other psychophysiological feature. Since heart beat interval seems to be the

only significant psychophysiological feature when using a CUSUM chart, we can take a sample with a fixed number of heartbeats. We choose to use sample sizes of one, three, five, and ten heartbeats. We are able to use a sample size of one since a CUSUM chart can be used with individuals (Montgomery, 1997). Using the fixed sample size data set, we recalculate the classification accuracies of each of the CUSUM charts using heart beat interval. The results from Pilot 1, Day 1 are in Table 4-2.

Table 4-2: False Alarm Rate, Misidentification Rates, and Classification Accuracies of Pilot 1, Day 1 Data Using a Fixed Sample Size CUSUM Chart

Sample Size	Value of H	False Alarm Rate	Misidentification Rate	Classification Accuracy
1	76.4	39.33%	0%	77.52%
3	21.5	39.28%	0%	77.57%
5	15.3	39.28%	0%	77.48%
10	7.7	39.26%	0%	77.59%

We found similar results using the other three data sets. These results are in Appendix E. The classification accuracies found using fixed sample size CUSUM charts are similar to those attained using variable sample size CUSUM charts. One thing that we did find out was that the classification accuracy is not sensitive to sample size.

Even though CUSUM charts with ceilings did not have the expected results with variable sample size CUSUM charts, using fixed sample sizes may improve this classification accuracy. The result from a CUSUM chart with a ceiling for pilot 1, day 1 is in Table 4-3.

Table 4-3: False Alarm Rate, Misidentification Rate, and Classification Accuracy of a CUSUM Chart with a Ceiling for Pilot 1, Day 1

Sample Size	Value of H	Ceiling	False Alarm Rate	Misidentification Rate	Classification Accuracy
1	76.4	21.1	33.64%	1.16%	80.28%
3	21.5	21.5	33.58%	0.10%	80.40%
5	15.3	22.0	33.40%	1.10%	80.42%
10	7.7	22.5	33.06%	1.10%	80.66%

We found that the classification accuracy did improve somewhat over traditional CUSUM charts. The addition of a ceiling improves the classification accuracy for pilot 1, but is insignificant for pilot 4. We found that the classification accuracy of a fixed sample size CUSUM chart is very similar to the previous classifiers. Therefore, it may be practical to use a CUSUM chart for individuals.

We also need to determine if fixed sample size CUSUM charts also improves the classification accuracies for different pilots and for different days. Since the sample size does not make a difference, we compare only the CUSUM chart for individuals. When we use each CUSUM classifier against other data sets, we find that the control charts maintain classification accuracies across different days. However, the classification accuracies decreased significantly when used to classify different pilots. The results from this comparison are in Table 4-4.

Table 4-4: Classification Accuracies Across Pilots for a CUSUM Chart for Individuals

Data Set	Classification Accuracy Against Pilot 1, Day 1	Classification Accuracy Against Pilot 1, Day 2	Classification Accuracy Against Pilot 4, Day 1	Classification Accuracy Against Pilot 4, Day 2
Pilot 1, Day 1	77.52%	71.38%	62.28%	73.37%
Pilot 1, Day 2	75.54%	76.80%	60.06%	72.52%
Pilot 4, Day 1	72.26%	59.90%	77.52%	73.60%
Pilot 4, Day 2	73.13%	59.90%	68.90%	77.20%

We also rebuilt each CUSUM chart with a ceiling and applied them across the four data sets. Using the modified CUSUM charts, we found similar results to the traditional CUSUM charts. The results are in Table 4-5.

Table 4-5: Classification Accuracies Across Pilots for a CUSUM Charts with Ceilings for Individuals

Data Set	Classification Accuracy Against Pilot 1, Day 1	Classification Accuracy Against Pilot 1, Day 2	Classification Accuracy Against Pilot 4, Day 1	Classification Accuracy Against Pilot 4, Day 2
Pilot 1, Day 1	80.28%	71.09	62.28%	73.37%
Pilot 1, Day 2	76.79%	79.18%	60.06%	72.53%
Pilot 4, Day 1	70.32%	59.90%	79.29%	73.74%
Pilot 4, Day 2	70.56%	59.90%	68.90%	77.35%

In both cases, using fixed sample size maintains the classification accuracies when used to classify pilot mental workload for the same day. The classification accuracy decreases when used to classify mental workload for other pilots

4.4. Summary

Modifications to the original control charts did not provide the improvement we had originally hoped. The CUSUM chart with a ceiling did not reduce the false alarm rate as originally thought. Changing the sample size of the original data set did not improve the classification accuracy. Often, these improvements actually decreased the classification accuracy of the control chart. Until we can change these modifications so that they improve the classification accuracy, traditional control charts are the best classifiers of pilot mental workload.

5. Conclusions

5.1. Effectiveness of Control Charts to Classify Pilot Workload

The purpose of this research is to determine if using control charts on different psychophysiological features can classify pilot mental workload. Although many of the control charts we built did not work, we did find a couple of control charts that did. Across these four data sets, we found that a CUSUM chart using heart beat interval as the sole factor attained the best classification accuracies out of the seven features. When using this control chart, we found that the classification accuracies are lower than those attained by East, as shown in Table 5-1.

Table 5-1: Comparison of SPC Classifiers Verses Multivariate Discriminant Classifiers Built by East for Same Pilot, Same Day Analysis

Data Set	SPC Classifier	Best Multivariate Classifier
Pilot 1, Day 1	78.79%	82.51%
Pilot 1, Day 2	87.40%	78.51%
Pilot 4, Day 1	77.57%	97.46%
Pilot 4, Day 2	77.65%	90.58%

Figure 5-1 shows how each of the four classifiers classified pilot workload across all twenty-two segments. It also shows the actual workload level as used by this research.

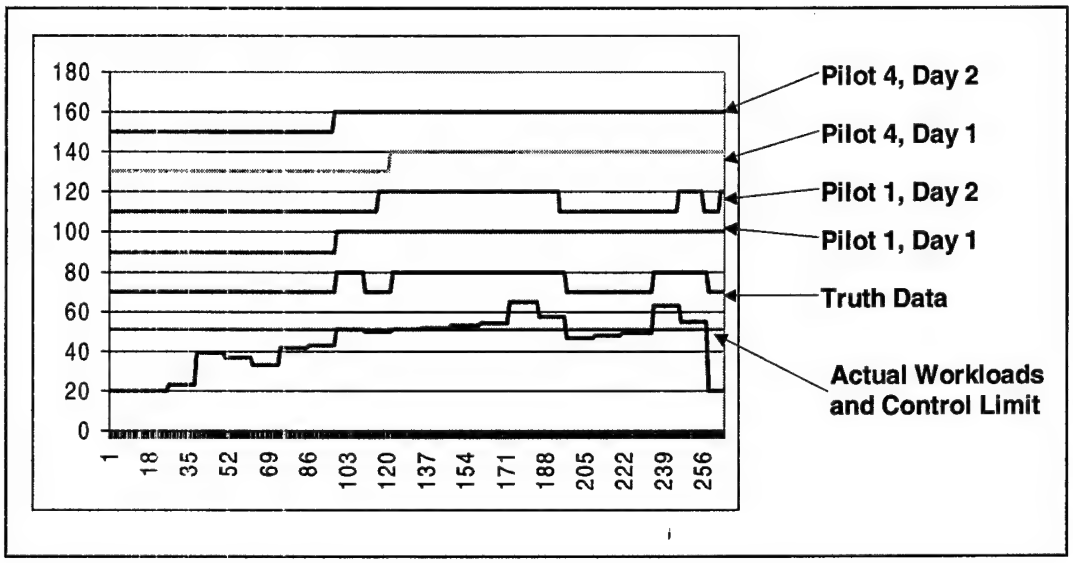

Figure 5-1: Classification of Pilot Workload for All Four Classifiers for All Segments

We find that each of the classifiers were accurate up until segment eight, where the workload shifts from low to high for the first time. Two of the classifiers were able to detect the shift in workload immediately, while the other two classifiers eventually detect the shift after a few minutes. Most of the errors occur after this initial shift, which indicates that the control charts are only able to detect the initial workload shift. This makes sense, as control charts are designed to detect a process that is out of control. In the case of pilot mental workload classification, the initial change from low to high workload is considered to be the first segment where the process is out of control. The control charts are effective because they are able to detect this initial change. However, because control charts are not designed to monitor processes after they are out of control, the accuracy of the control chart decreases significantly after the initial shift to high workload.

The classifiers that we created did not have the classification accuracies as previous classifiers. However, there are two advantages to the SPC classifier. The SPC classifier is much simpler to implement, as it tracks only one feature, heart beat interval. It may be simpler and more cost effective to track only one feature. The second advantage the SPC classifier has over

multivariate discriminant classifiers is that control charts are robust enough to maintain classification accuracies across pilots and across days. Table 5-2 shows the comparisons of both classifiers across different pilots and different days.

Table 5-2: Comparison of SPC and Multivariate Classifiers Across Different Pilots and Days

Data Set	Classifier Type	Classification Accuracy Against Pilot 1, Day 1	Classification Accuracy Against Pilot 1, Day 2	Classification Accuracy Against Pilot 4, Day 1	Classification Accuracy Against Pilot 4, Day 2
Pilot 1 Day 1	SPC	78.79%	87.40%	75.29%	75.39%
	Multivariate	82.51%	63.24%	47.43%	66.80%
Pilot 1 Day 2	SPC	78.79%	87.40%	75.29%	75.39%
	Multivariate	64.43%	78.51%	48.22%	72.02%
Pilot 4 Day 1	SPC	78.03%	83.97%	77.57%	74.24%
	Multivariate	59.09%	59.09%	97.46%	60.87%
Pilot 4 Day 2	SPC	57.95%	77.60%	68.82%	77.65%
	Multivariate	60.87%	61.86%	53.16%	90.58%

Control charts have better classification accuracies across different pilots and days because we can use data set specific means and standard deviations as part of the classifier. We found that using multiple control charts did not increase the accuracy of the classifier as we had hoped. On the contrary, using multiple control chart greatly decrease the classification accuracy of the classifier. Even the classifiers that use a combination of the most accurate control charts did not improve the classification accuracy of this classifier. Therefore, single-feature control charts are currently the best type of classifier we found from this research.

5.2. Recommendations

5.2.1. Use Different Psychophysiological Features

The only psychophysiological features used in this research were cardiac, ocular, and respiratory measurements. We chose these features because they were found to be significant in past research. The other features that were found to be significant in past research were brain activity measurements. Since we know that control charts can classify pilot mental workload

using these three categories of features, it may be beneficial to use a control chart with brain activity measurements, which were found to be significant from past research. Past research has also created different factors, which are linear combinations of physiological features. We may find that a control chart using one of these factors may produce even higher classification accuracies.

5.2.2. Reclassify the Data Set Workload

One of the assumptions made during this research is that all 22 segments are classified correctly. It is possible that some of these segments are misclassified. This misclassification could account for some of the unexplained inaccuracies of the model, especially the high occurrence of false alarm rates found in the CUSUM chart classifiers.

Another explanation is that the pilot may be in a period that is classified as low mental workload, but his psychophysiological features are consistently showing that he actually in a period of high mental workload. These misclassifications may occur for many reasons. One instance is when the pilot may start to show signs of high mental workload during the period prior to actually entering high workload. This can be attributed to anxiety or nervousness a pilot may experience before a mentally demanding task. Also, the pilot may know of an upcoming task, and mentally prepare for it. When this happens, the pilot shows signs of being in high mental workload before entering a period of high mental stress. A similar type of misclassification occurs during the segment right after completing a mentally challenging task. This is because the psychophysiological features need time to return to a period of low workload, even though the stressor may be gone. It is possible that the pilot's features never return to normal during short low workload segments, due to the cool down period from the previous high workload segment, and then the ramp up to the next high workload segment. Most of the false alarms occurred during these warm-up and cool-down periods for different control charts and data sets. The high occurrence of these false alarms at the same time intervals suggests that these segments can be reclassified.

To test this theory, we rebuilt the data set by classifying the minute before and after each high workload segment as high instead of low. This accounts for ramp-up and cool-down periods where the pilot may still exhibit signs of high mental workload. We rebuilt the CUSUM chart using heart beat interval for pilot 1, day 1 and optimize the values of k based upon the classification accuracies. We found a substantial increase in the classification accuracies. The classification accuracy of using heart beat interval CUSUM chart (the best classifier we found using the original workload levels) increased from 78.79 to 90.2%.

We apply the new workload classification to all four data sets. Using a CUSUM chart with heart beat interval, the classification accuracies for all four data sets were above 85%. The results from this are in Table 5-3.

Table 5-3: False Alarm Rates, Misidentification Rates, and Classification Accuracies of Heart Beat Interval for Pilots 1 and 4 for Both Days

Pilot Number and Day	False Alarm Rate	Misidentification of High Workload Rate	Classification Accuracy
Pilot 1, Day 1	22.2%	0%	90.2%
Pilot 1, Day 2	22.2%	6.1%	86.7%
Pilot 4, Day 1	23.1%	0%	89.8%
Pilot 4, Day 2	22.2%	0%	90.2%

These results show that the reclassified data set may be a more accurate representation of pilot workload. We apply the four different classifiers against the other data sets to determine the classification accuracy across pilots and across days. We found that the classification accuracy across pilots and days did not diminish at all. The results from this analysis is in Table 5-4.

Table 5-4: Classification Accuracy Across Pilots and Days Using Heart Beat Interval as the Primary Classifier

Classifier	Pilot 1, Day 1	Pilot 1, Day 2	Pilot 4, Day 1	Pilot 4, Day 2
Pilot 1, Day 1 Heart Beat Interval	90.2%	84.1%	83.3%	84.1%
Pilot 1, Day 2 Heart Beat Interval	71.6%	86.7%	83.3%	88.3%
Pilot 2, Day 1 Heart Beat Interval	75.0%	86.4%	89.8%	87.1%
Pilot 2, Day 2 Heart Beat Interval	70.8%	81.1%	74.2%	90.2%

5.2.3. Using a Two-Hypothesis Test Control Chart

The original hypothesis for a control chart is that "a process is in-control unless there is evidence to prove otherwise." In the case of pilot mental workload, that hypothesis translates to "a pilot is in a state of low workload, unless there is evidence to prove he is in a state of high mental workload." Every control chart built during this research effort uses the above hypothesis, which s applied to every data sample, regardless of workload level. The addition of a second hypothesis may make future control charts more accurate. The second hypothesis would be "a process is out-of-control unless there is evidence to prove otherwise." In the case of pilot mental workload, that hypothesis translates to "a pilot is in a state of high workload, unless there is evidence to prove he or she is in a state of low mental workload."

The choice of which hypothesis to use depends upon the current state of the pilot, as determined by the control chart. If the pilot is in a state of low mental workload, then the first hypothesis becomes the null hypothesis. A rejection of the null hypothesis test changes the mental workload from low to high. Then, the new null hypothesis changes to the second hypothesis. The change to the new hypothesis may require the construction of a new control chart using a different mean and standard deviation. Figure 5-2 is an example of how this new type of control chart might look.

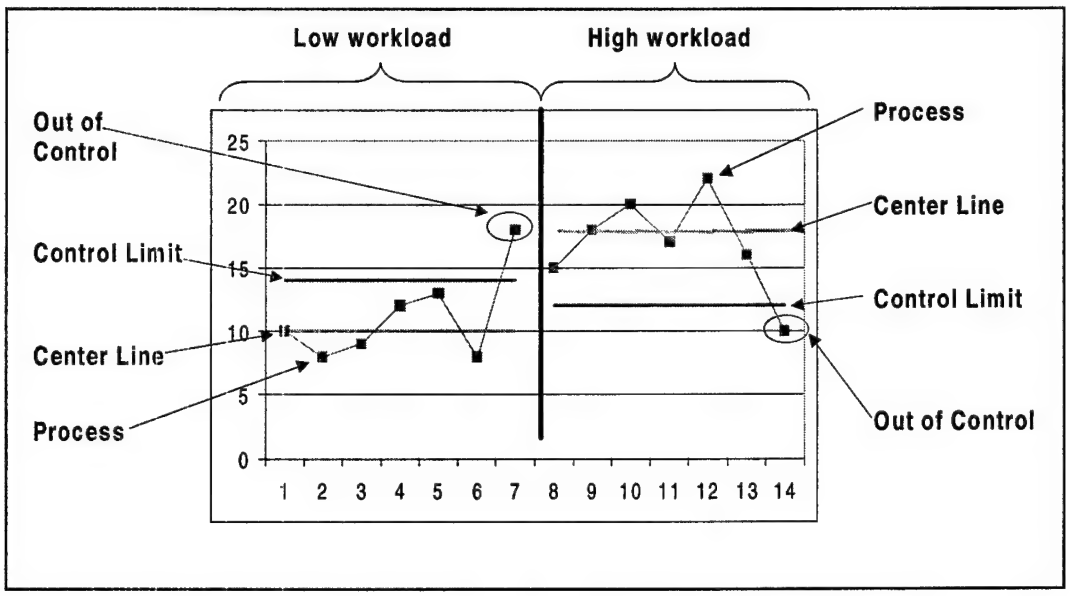

Figure 5-2: Example of a Two-Hypothesis Test Control Chart

The difference between this type of control chart and a normal control chart is the method used to classify high workload. The control charts we built during this research watched for evidence of a shift from low to high workload only. The modified control chart would require that there be enough evidence to show that the pilot is in a state of low mental workload before resuming that hypothesis. Such a system should reduce the false alarm rate following an initial shift to high workload.

Appendix A

False Alarm Rates, Misidentification Rates, and Classification Accuracies Based Upon Single Feature X-bar Charts with Three-Sigma Control Limits

Data for Pilot 1, Day 1

Feature	False Alarm Rate	Mis-identification of High Workload Rate	Overall Classification Accuracy
Breath Interval	3.3%	92.8%	59.1%
Breath Min Amp	13.7%	90.1%	54.2%
Breath Max Amp	3.9%	95.5%	57.6%
Blink Interval	15.7%	65.8%	63.3%
Blink Amplitude	5.2%	91.0%	58.7%
Blink Duration	5.2%	92.8%	58.0%

Data for Pilot 1, Day 2

Feature	False Alarm Rate	Mis-identification of High Workload Rate	Overall Classification Accuracy
Breath Interval	5.9%	92.8%	57.6%
Breath Min Amp	0%	100%	58.0%
Breath Max Amp	4.6%	94.6%	57.6%
Blink Interval	11.1%	78.4%	60.7%
Blink Amplitude	0%	100%	58.0%
Blink Duration	2.6%	96.4%	58.0%

Data for Pilot 4, Day 1

Feature	False Alarm Rate	Mis-identification of High Workload Rate	Overall Classification Accuracy
Breath Interval	13.1%	94.6%	42.7%
Breath Min Amp	0%	100%	58.0%
Breath Max Amp	14.3%	89.2%	54.2%
Blink Interval	15.7%	88.3%	53.8%
Blink Amplitude	1.3%	100%	57.2%
Blink Duration	0%	100%	58.0%

Data for Pilot 4, Day 2

Feature	False Alarm Rate	Mis-identification of High Workload Rate	Overall Classification Accuracy
Breath Interval	5.2%	94.6%	57.2%
Breath Min Amp	0%	100%	58.0%
Breath Max Amp	0%	100%	58.0%
Blink Interval	7.2%	87.6%	59.1%
Blink Amplitude	0%	99.1%	58.0%
Blink Duration	1.3%	100%	57.2%

Appendix B:

False Alarm Rates, Misidentification Rates, and Classification Accuracies for Pilot 1, Day 1

Based Upon Single Feature X-bar Charts with Variable Sigma Control Limits

Pilot 1, Day 1

Feature	Value of k	False Alarm Rate	Misidentification of High Workload Rate	Overall Classification Accuracy
Breath Interval	1.6	3.27%	84.68%	62.50%
Breath Min Amp	20.0	11.76%	100%	51.14%
Breath Max Amp	4.0	29.41%	65.77%	55.30%
Blink Interval	5.0	16.99%	74.77%	58.71%
Blink Amplitude	9.0	30.72%	18.92%	74.20%
Blink Duration	2.5	38.56%	9.01%	73.86%

Pilot 1, Day 2

Feature	Value of k	False Alarm Rate	Misidentification of High Workload Rate	Overall Classification Accuracy
Breath Interval	0.7	17.88%	67.27%	61.30%
Breath Min Amp	2.0	5.96%	100%	54.41%
Breath Max Amp	4.0	5.96%	93.64%	57.09%
Blink Interval	4.0	10.60%	78.18%	60.92%
Blink Amplitude	3.0	29.14%	49.09%	62.45%
Blink Duration	1.0	37.75%	22.73%	68.58%

Pilot 4, Day 1

Feature	Value of k	False Alarm Rate	Misidentification of High Workload Rate	Overall Classification Accuracy
Breath Interval	6.0	10.97%	94.44%	54.75%
Breath Min Amp	2.0	21.29%	100%	46.39%
Breath Max Amp	20.0	21.29%	86.11%	52.09%
Blink Interval	13.0	6.45%	93.52%	57.79%
Blink Amplitude	4.0	15.48%	86.11%	53.51%
Blink Duration	1.3	16.77%	35.19%	75.67%

Pilot 4, Day 2

Feature	Value of k	False Alarm Rate	Misidentification of High Workload Rate	Overall Classification Accuracy
Breath Interval	3.0	5.16%	94.50%	57.95%
Breath Min Amp	4.0	5.81%	100%	55.30%
Breath Max Amp	33.0	13.55%	82.57%	57.95%
Blink Interval	2.0	7.74%	82.57%	61.36%
Blink Amplitude	3.0	10.97%	86.24%	57.95%
Blink Duration	1.3	17.42%	30.28%	77.27%

Appendix C:

False Alarm Rates, Misidentification Rates, and Classification Accuracies Based Upon Number of Charts Signaling

Pilot 1, Day 1

Number of Charts to Produce a Signal	False Alarm Rate	Misidentification of High Workload Rate	Classification Accuracy
1 of 7 signals	45.10%	66.67%	45.83%
2 of 7 signals	14.38%	86.49%	55.30%
3 of 7 signals	12.42%	89.19%	55.30%
4 of 7 signals	11.76%	96.40%	52.65%
5 of 7 signals	7.19%	100%	53.79%
6 of 7 signals	1.96%	100%	56.82%
7 of 7 signals	0%	100%	57.95%

Pilot 1, Day 2

Number of Charts to Produce a Signal	False Alarm Rate	Misidentification of High Workload Rate	Classification Accuracy
1 of 7 signals	32.24%	79.07%	48.09%
2 of 7 signals	7.24%	93.64%	56.49%
3 of 7 signals	5.92%	94.55%	56.87%
4 of 7 signals	5.92%	96.36%	56.11%
5 of 7 signals	1.32%	100%	57.25%
6 of 7 signals	0%	100%	58.02%
7 of 7 signals	0%	100%	58.02%

Pilot 4, Day 1

Number of Charts to Produce a Signal	False Alarm Rate	Misidentification of High Workload Rate	Classification Accuracy
1 of 7 signals	39.35%	79.63%	44.11%
2 of 7 signals	23.23%	86.11%	50.95%
3 of 7 signals	21.29%	86.11%	52.09%
4 of 7 signals	21.29%	87.04%	51.71%
5 of 7 signals	12.23%	95.37%	53.61%
6 of 7 signals	1.94%	99.07%	58.17%
7 of 7 signals	0%	100%	58.94%

Pilot 4, Day 2

Number of Charts to Produce a Signal	False Alarm Rate	Misidentification of High Workload Rate	Classification Accuracy
1 of 7 signals	34.84%	76.15%	48.10%
2 of 7 signals	5.81%	95.41%	57.20%
3 of 7 signals	5.81%	96.33%	56.82%
4 of 7 signals	5.81%	97.25%	56.44%
5 of 7 signals	3.87%	99.08%	56.82%
6 of 7 signals	0%	100%	58.71%
7 of 7 signals	0%	100%	58.71%

Appendix D:

Results Using a Traditional CUSUM Chart

Pilot 1, Day 1

Feature	Value of H	False Alarm Rate	Misidentification Rate	Classification Accuracy
Heart Rate	53	36.60%	0%	78.79%
Breath Interval	0.3	3.92%	83.78%	62.50%
Breath Min Amp	40	0%	100%	58.02%
Breath Max Amp	70	0%	100%	58.02%
Blink Interval	3.1	12.42%	75.68%	60.98%
Blink Amplitude	350	0%	100%	58.02%
Blink Duration	50	0%	100%	58.02%

Pilot 1, Day 2

Feature	Value of H	False Alarm Rate	Misidentification Rate	Classification Accuracy
Heart Rate	53	3.95%	24.55%	87.40%
Breath Interval	2	0%	100%	57.95%
Breath Min Amp	3	0%	100%	57.95%
Breath Max Amp	20	0%	100%	57.95%
Blink Interval	2	1.97%	90.91%	60.59%
Blink Amplitude	3	0%	100%	57.95%
Blink Duration	80	0%	100%	57.95%

Pilot 4, Day 1

Feature	Value of H	False Alarm Rate	Misidentification Rate	Classification Accuracy
Heart Rate	48	38.06%	0%	77.57%
Breath Interval	5	0%	100%	58.94%
Breath Min Amp	30	0%	100%	58.94%
Breath Max Amp	10	0%	100%	58.94%
Blink Interval	3	0%	100%	58.94%
Blink Amplitude	10	0%	100%	58.94%
Blink Duration	2	0%	100%	58.94%

Pilot 4, Day 2

Feature	Value of H	False Alarm Rate	Misidentification Rate	Classification Accuracy
Heart Rate	17	38.06%	0%	77.65%
Breath Interval	4	0%	100%	58.71%
Breath Min Amp	2	0%	100%	58.71%
Breath Max Amp	2	0%	100%	58.71%
Blink Interval	3	0%	100%	58.71%
Blink Amplitude	3	0%	100%	58.71%
Blink Duration	3	0%	100%	58.71%

Appendix E:

CUSUM Charts Using Heart Beat Interval with Fixed Sample Sizes

Traditional CUSUM Chart for Pilot 1, Day 1

Sample Size	Value of H	False Alarm Rate	Misidentification Rate	Classification Accuracy
1	76.4	39.33%	0%	77.52%
3	21.5	39.28%	0%	77.57%
5	15.3	39.28%	0%	77.48%
10	7.7	39.26%	0%	77.59%

CUSUM Chart with Ceiling for Pilot 1, Day 1

Sample Size	Value of H	Ceiling	False Alarm Rate	Misidentification Rate	Classification Accuracy
1	76.4	21.1	33.64%	1.16%	80.28%
3	21.5	21.5	33.58%	0.10%	80.40%
5	15.3	22.0	33.40%	1.10%	80.42%
10	7.7	22.5	33.06%	1.10%	80.66%

Traditional CUSUM Chart for Pilot 1, Day 2

Sample Size	Value of H	False Alarm Rate	Misidentification Rate	Classification Accuracy
1	64.9	30.83%	12.71%	76.80%
3	21.6	30.80%	12.78%	76.80%
5	13.0	30.74%	12.62%	76.91%
10	6.5	31.05%	13.37%	76.41%

CUSUM Chart with Ceiling for Pilot 1, Day 2

Sample Size	Value of H	Ceiling	False Alarm Rate	Misidentification Rate	Classification Accuracy
1	64.9	35.5	26.57%	12.91%	79.18%
3	21.6	36.0	26.57%	12.93%	79.19%
5	13.0	38.0	26.58%	12.62%	79.31%
10	6.5	39.0	27.08%	13.37%	78.71%

Traditional CUSUM Chart for Pilot 4, Day 1

Sample Size	Value of H	False Alarm Rate	Misidentification Rate	Classification Accuracy
1	190.5	39.74%	0%	77.52%
3	63.6	39.70%	0%	77.54%
5	38.2	39.72%	0%	77.55%
10	19.1	39.55%	0%	77.71%

CUSUM Chart with Ceiling for Pilot 4, Day 1

Sample Size	Value of H	Ceiling	False Alarm Rate	Misidentification Rate	Classification Accuracy
1	190.5	0.1	36.61%	0%	79.29%
3	63.6	0.1	36.66%	0%	79.26%
5	38.2	0.1	36.76%	0%	79.22%
10	19.1	0.1	36.72%	0%	79.30%

Traditional CUSUM Chart for Pilot 4, Day 2

Sample Size	Value of H	False Alarm Rate	Misidentification Rate	Classification Accuracy
1	125.5	41.74%	0%	77.20%
3	42	41.72%	0%	77.22%
5	25.4	41.62%	0%	77.31%
10	12.8	41.52%	0%	77.40%

CUSUM Chart with Ceiling for Pilot 4, Day 2

Sample Size	Value of H	Ceiling	False Alarm Rate	Misidentification Rate	Classification Accuracy
1	125.5	0.01	41.46%	0%	77.35%
3	42	0.01	41.41%	0%	77.39%
5	25.4	0.01	41.45%	0%	77.40%
10	12.8	0.01	41.52%	0%	77.40%

Bibliography

Air Force Research Laboratory, AFRL, "Flight Psychophysiology Laboratory," 1998, Office Brochure, Flight Psychophysiology Laboratory, Human Interface Technology Branch, Crew System Interface Division, Human Effectiveness Directorate (AFRL/HECP).

Alwan, L.C. and Roberts, H.V. 1988 "Time-series Modeling for Statistical Process Control." Journal of Business & Economic Statistics, 6(1): pp. 87-95.

Annadi, H.P., Keats, J.B., Runger, G.C., and Montgomery, D.C. "An Adaptive Sample Size CUSUM Control Chart." International Journal of Production Research, Vol 33, No 6. pp. 1605-1616, 1995.

Auten, J. 1996 "G-LOC: Is the Cluebag Half Full or Half Empty?" Flying Safety, Vol 52, pp.5-6

Box, G. and Kramer, T. "Statistical Process Monitoring and Feedback adjustment-A Discussion," Technometrics, 34(3): 251-285 (1992).

Box, G.E., Jenkins, G.M., and Reinsel, G.C. Time Series Analysis, (Third Edition). Prentice-Hall Inc, 1994.

Damos, D.L., editor. *Multiple-tasks Performance.* London: Taylor and Francis Ltd., 1996. 328-360.

East, J.A., Feature Selection for Predicting Pilot Mental Workload, M.S. Thesis, Air Force Institute of Technology, Wright-Patterson AFB OH.

Greene, K.A. (1998) Feature Saliency in Artificial Neural Networks with Application to Modeling Workload, Ph.D Dissertation, Air Force Institute of Technology, Wright-Patterson AFB OH.

Hankins, T.C. and Wilson, G.F."A Comparison of Heart Rate, Eye Activity, EEG and subjective Measures of Pilot Mental Workload During Flight," Aviation Space Environmental Medicine 69:360-367 (1998).

Laine, T.I. (1999) Selection of Psychophysiological Features Across Subjects for Classifying Workload Using Artificial Neural Networks, M.S. Thesis, Air Force Institute of Technology, Wright-Patterson AFB OH.

Liu, W.C. (1997) The Application of Statistical Process Control to Departure Reliability Improvement, M.S. Thesis, Air Force Institute of Technology, Wright-Patterson AFB OH.

Mastrangelo, C.M. and Montgomery, D.C. "Some Statistical Process Control Methods For Autocorrelated Data," Journal of Quality Technology, 23(3):179-193 (July 1991).

Montgomery, D.C. Introduction to Statistical Quality Control, (Third Edition). John Wiley & Sons, 1997.

Reynolds, M.R. Jr., Amin, R.W. and Arnold, J.C., 1991, "CUSUM Charts with Variable Sampling Intervals," Technometrics, 32, 371-384.

Richardson, B.D. (1996) Statistical Process control and Medical Surveillance, M.S. Thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH.

Wilson, G.F. "Applied Use of Cardiac and Respiration Measures: Practical Considerations and Precautions," *Biological Psychology,* 34:163-178 (1992).

Wilson, G.F. "Air-to-Ground Training Missions: A Psychophysiological Workload Analysis," *Ergonomics,* 36(9):1071-1087 (1993).

Wilson, G.F. and Fisher, F. "Cognitive Task Classification Based Upon Topographical EEG Data," *Biological Psychology,* 40:239-250 (1995).

Wilson, G.F. and Fisher, F. "The Use of Cardiac and Eye Blink Measures to Determine Flight Segment in F4 Crews," *Aviation, Space, and Environmental Medicine,* 33:959-962 (October 1997).

Wilson, G.F., et al. "Evoked Potential, Cardiac, Blink, and Respiration Measures of Pilot Workload in Air-to-Ground Missions," *Aviation, Space, and Environmental Medicine.*

Young, R.R. (1998) Multivariate Analysis and Statistical Process Control of Steering Wheel Manufacturing Deviation Data, M.S. Thesis, Air Force Institute of Technology, Wright-Patterson AFB OH.

Zalewski, D.J. (1995) Methods for Monitoring Process control and Capability in the Presence of Autocorrelation, Ph. D Dissertation, Air Force Institute of Technology, Wright-Patterson AFB, OH.

Vita

Captain Terence Y. Kudo was born in Honolulu, Hawaii. He graduated from Iolani High school in 1992, and entered the United States Air Force Academy as a member of the class of 1996. He graduated with a Bachelor of Science degree in Operations Research in June of 1996.

His first assignment was at Wright-Patterson AFB at the National Air Intelligence Center. He worked there for three years until he entered the Graduate School of Engineering and Management, Air Force Institute of Technology in August 1999. Upon graduation, he will be assigned to the 422TES at Nellis AFB, NV.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
20-03-2001	Master's Thesis	Jun 2000 – Sep 2001

4. TITLE AND SUBTITLE

USING STATISTICAL PROCESS CONTROL METHODS TO CLASSIFY PILOT MENTAL WORKLOAD

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kudo, Terence, Y., Captain, USAF

5d. PROJECT NUMBER
If funded, enter ENR #

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 P Street, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GOR/ENS/00M-10

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Glenn F. Wilson, PhD. AFRL/HECP
2255 H. Street, Bldg 33
Wright-Patterson AFB, OH 45433-7022
(937)-785-8748 Glenn. Wilson@wpafb.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The problem of classifying pilot mental workload is important to the United States Air Force. Pilots are more subject to errors and G-induced loss of consciousness during periods of mental overload and task saturation. Often the result is the loss of aircraft, and in extreme cases, the loss of the pilot's life. Current research efforts use different psychophysiological featuers to classify pilot mental workload. These include cardiac, ocular, respiratory, and brain activity measures.

The focus of this effort is to apply statistical process control methodology on different psychophysiological features in an attempt to classify pilot mental workload. The control charts track these features throughout the flight, and classify a segment as high workload if the measurements of these features are greater than predefined control limits. We find that using certain control charts prove to be effective workload classifiers and maintain high classification accuracies when applied to other flight data.

15. SUBJECT TERMS

Statistical Process Control, Quality Control, Control Charts, Classifiers
Pilot Mental Workload, Mental Stress, Psychophysiological Features

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	b. ABSTRACT	c. THIS PAGE			Jeffrey W. Lanning, Major, USAF
U	U	U	UU	90	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-6565 x4324

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18

Form Approved
OMB No. 074-0188